40

ajer

THE ALBERTA JOURNAL OF EDUCATIONAL RESEARCH

VOLUME XL NUMBER 2 JUNE 1994

PUBLISHED BY
THE UNIVERSITY OF ALBERTA • EDMONTON

The Alberta Journal of Educational Research

Published in March, June, September, and December
by the Faculty of Education, University of Alberta

AJER is a quarterly journal devoted to the dissemination, criticism, interpretation, and encouragement of all forms of systematic enquiry into education and fields related to or associated with education.

Editor: Judy Cameron

Managing Editor: Naomi Stinson
Proofreader: Karen McFarlane

Consulting Editors

Terry Belke
University of Alberta

George Buck
University of Alberta

James Chapman
Massey University

John Connors
Canadian Union College

Samuel Deitz
Georgia State University

Sharon M. Haggerty
University of Western Ontario

Linda LaRocque
University of Alberta

Stewart J.H. McCann
University College of Cape Breton

Antoinette Oberg
University of Victoria

Ruth Rees
Queen's University

Sharon Robertson
University of Calgary

Robert H. Short
University of Alberta

Kelleen Toohey
Simon Fraser University

Robert Wilson
Queen's University

Faculty Advisory Committee

John G. Paterson (Chair)
Joyce Edwards
Graham Fishburne
Robert H. Short
Fern Snart

AJER gratefully acknowledges support from the Social Sciences and Humanities Research Council of Canada and from the Alberta Advisory Committee for Educational Studies.

The subscription rate is $32.00 per year for individuals, $40.00 per year for institutions. Add $8.00 for delivery outside Canada. Single copies are $8.00 each. Subscriptions and sales in Canada will be charged 7% GST. Please make cheques payable to *The Alberta Journal of Educational Research*. Back issues are available; rates supplied on request. Claims for undelivered copies must be received within three months of publication.

Address all communications and manuscript submissions to:

The Alberta Journal of Educational Research
Publication Services
4-116 Education Centre North
University of Alberta
Edmonton, AB, Canada, T6G 2G5.
Fax: (403) 492-0390

Publications Mail Registration Number 1436

Faculty of Education
University of Alberta

The Alberta Journal of Educational Research

Volume XL, Number 2, June 1994

Robert J. Wilson *John R. Kirby*	**105**	Introduction: Special Issue on Cognition and Assessment
Thomas Maguire *John Hattie* *Brian Haig*	**109**	Construct Validity and Achievement Assessment
Gay L. Bisanz *Jeffrey Bisanz*	**127**	Studying the Development of Academic Skills in the 1990s: Implications for Assessment
John R. Kirby *Rosamund A. Woodhouse*	**147**	Measuring and Predicting Depth of Processing in Learning
Philip Nagy	**163**	Cognitive Assessment of Solutions to an Ill-structured Problem Using Forced Classification Analysis
Philip H. Winne *Lorene Gupta* *John C. Nesbit*	**177**	Exploring Individual Differences in Studying Strategies Using Graph Theoretic Statistics
W. Todd Rogers *David J. Bateson*	**195**	Verification of a Model of Test-taking Behavior of High School Seniors
Bikkar S. Randhawa	**213**	Theory, Research, and Assessment of Mathematical Problem Solving
David Bateson	**233**	Psychometric and Philosophic Problems in "Authentic" Assessment: Performance Tasks and Portfolios
Dan G. Bachor *John O. Anderson* *John Walsh* *Walter Muir*	**247**	Classroom Assessment and the Relationship to Representativeness, Accuracy, and Consistency

ISSN 0002-4805

The Alberta Journal of Educational Research *Vol. XL, No. 2, June 1994, 105-108*

Robert J. Wilson

and

John R. Kirby
Queen's University

Introduction: Special Issue on Cognition and Assessment

The educational field of measurement has changed dramatically in recent years. Teachers are unenthusiastic about the value of large-scale, centrally designed testing, perceiving it to be less than relevant to classroom learning. Administrators and parents worry about some of the methods teachers have adopted for assessing students' success. As the wider community too has begun to challenge the effectiveness of the education system in general, concerns with measuring learning have taken on a political urgency. The time is ripe to reconsider the goals and techniques of educational assessment.

In measurement terms, these issues concern validity, but validity in measuring achievement outcomes is becoming more elusive even as more attention is being paid to it. As some investigators concentrate on the contexts in which assessment takes place and the representativeness of tasks used in those contexts, others concern themselves with the adequacy of any tasks to measure cognitive operations assumed to be the mental resources that create achievement outcomes.

Central to all these interests are the processes invoked to respond to achievement tasks. For some measurement specialists, Bloom's Taxonomy has been the beginning and end of the consideration of processes, although the taxonomy itself is dated and superseded by much recent work in cognition. Conversely, some work in cognition has often treated validity concerns such as construct span and representativeness inadequately and disinterestedly. Both cognition and measurement stand to benefit from a consideration of each other's problems and solutions (Baker, O'Neil & Linn, 1993; Wittrock & Baker, 1991).

Two previous meetings of Canadian researchers interested in measurement issues (Anderson, 1990; Bateson, 1992) led the participants into a concern with the nature of achievement, cognitive processes, and the ways these are assessed. A third meeting, it was decided, should exploit Cognition and Assessment as a theme, and papers were invited from researchers active in both fields. This special issue of the *Alberta Journal of Educational Research* is a result of their efforts.

Robert Wilson is a professor of measurement and evaluation. His research interests are in large-scale assessment and classroom processes in evaluating student learning.
John Kirby is a professor of educational psychology. His research interests include individual differences in cognitive processes, learning disabilities, and study processes.

Measurement and Psychology

The fields of measurement and psychology have been closely allied throughout their histories. In many universities they inhabit the same department. Because educational measurement has been concerned primarily with assessing students' knowledge, skills, and problem solving; and cognitive psychology has been concerned with how knowledge, skills, and problem solving function are represented in the brain, it has seemed sensible to consider them as mutually supporting and enriching.

Until recently educational measurement has been closely aligned with psychometrics, a tradition in psychology that is more mathematical than cognitive. Although psychometrics (literally measurement of the mind) should be cognitive in orientation, it began its development during the behaviorist domination of learning theory. Lacking guidance on appropriate cognitive constructs to pursue, psychometricians focused instead on the development of mathematical techniques such as factor analysis for analyzing test results, thus shifting attention away from the need to conceptualize about the cognitive nature of the mathematical constructs. The result was a technology of assessment that grew out of a test theory that lacked a basis in psychological theory, but that nevertheless came to be a distinct scientific discipline in education and psychology (Cronbach, 1957; Ferguson, 1954).

Cognitive psychology should be a source for the constructs that underlie educational achievement, and may be able to shed some light on how to measure these constructs (Biggs & Collis, 1982). However, cognitive psychology had its roots in experimental psychology, the domain of operational definitions. To illustrate: when two experimental conditions are contrasted for their effect on concept identification, great care is taken in constructing the stimulus lists, but little thought is devoted to the representativeness of the content, the instructional method, or the particular measure of concept learning.

Validity issues are inescapable, but cognitive psychologists seem on safe ground in such studies—until they begin to generalize. Once inferences are drawn from this work about how to teach students real concepts, the concern for validity grows. The issue becomes even more important when cognitive psychologists turn their attention to more complex constructs such as those in theories of mental ability (e.g., Carroll, 1976; Das, Kirby, & Jarman, 1975; Sternberg, 1977). Although it is unlikely that cognitive psychology has any simple solutions for educational assessment, it may be able to contribute to finding those solutions through its work on modeling underlying cognitive processes that give rise to observed measurements.

In this issue we see a number of cognitive psychologists approaching the assessment of educational achievement, employing the techniques of cognitive psychology, but with concern for validity: Bisanz and Bisanz address the cognitive processes involved in reading and arithmetic; Kirby and Woodhouse consider methods for predicting and measuring depth of processing in learning; Nagy presents a technique for analyzing solutions to ill-structured problems that might help describe learners' cognitive processes; Randhawa focuses on the processes of mathematical problem solving; and Winne, Gupta, and Nesbit illustrate techniques for measuring students' learning strategies.

Authentic Assessment

Those who work in educational assessment did not need cognitive psychology to begin to change. There has been growing disenchantment among educators over traditional forms of educational assessment. Among other concerns has been the sense that centrally planned, standard-format tests do not tap what teachers are trying to teach, or even what curricula indicate they should be teaching. New assessment methods have been devised under the collective label "authentic assessment" (Wiggins, 1989). Bateson in this issue provides a useful overview of how this movement has both annoyed and inspired measurement practitioners, especially in the field of large-scale assessment.

Validity (of which "authenticity" is but one aspect) is specifically addressed in this issue in the articles by Maguire, Hattie, and Haig, and Bachor, Anderson, Walsh, and Muir. Among other issues, Maguire et al. discuss whether a crucial aspect of a measure's validity is the use to which the information is put. They emphasize the need for conceptual links between the construct and its indicator on the one hand, and the indicator and any score on the other. Bachor et al. are also interested in the use of measures. Their approach, however, is to delineate which aspects of tasks are more or less useful for common assessment needs in classrooms. Both of these articles expand the context in which the validity of educational assessment is considered.

Final Words

The articles in this special issue revolve around one central issue: how can cognitive psychology and measurement practice work together to enhance the valid assessment of learning? We believe that these articles contribute to a deepening of the debate about the purposes and methods of educational assessment and provide direction for future work.

We begin with the Maguire, Hattie, and Haig article which raises many of the fundamental issues regarding validity and assessment. The second article, by Bisanz and Bisanz, reviews recent research in cognitive psychology that is directly relevant to the assessment of children's learning of arithmetic and reading. Kirby and Woodhouse focus on the later skills of learning from text, and how these might be assessed. The following articles, by Nagy, Winne et al., Rogers and Bateson, and Randhawa, report a series of empirical studies that address the measurement of achievement in cognitive frameworks. The final articles, by Bateson and Bachor et al., revisit the issue of validity and return the discussion to the implications for classroom practice.

Acknowledgments

The conference on which this special issue is based was supported by the Social Sciences and Humanities Research Council of Canada, the School of Graduate Studies and Research and the Faculty of Education of Queen's University, whose help is gratefully acknowledged. We would like to thank Lyn Shulha for her help in organizing the conference, and Gay Bisanz, Phil Nagy, Randy Randhawa, and Phil Winne for their helpful comments on this introduction.

References

Anderson, J.A. (Ed.). (1990). Assessment in the classroom. Special Issue: *Alberta Journal of Educational Research, 36.*

Baker, E.L., O'Neil, H.F. Jr., & Linn, R.L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48,* 1210-1218.

Bateson, D.J. (1992). *Classroom testing in Canada.* Vancouver: Centre for Applied Studies in Evaluation (CASE), UBC.

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning: The SOLO Taxonomy*. New York: Academic Press.

Carroll, J.B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect." In L.B. Resnick (Ed.), *The nature of intelligence* (pp. 27-56). Hillsdale, NJ: Erlbaum.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671-684.

Das, J.P., Kirby, J.R. & Jarman, R.F. (1975). Simultaneous and successive synthesis: An alternative model for cognitive abilities. *Psychological Bulletin, 82*, 87-103.

Ferguson, G.A. (1954). On learning and human ability. *Canadian Journal of Psychology, 8*, 95-112.

Sternberg, R.J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

Wittrock, M.C., & Baker, E.L. (1991). (Eds.). *Testing and cognition*. Englewood Cliffs, NJ: Prentice-Hall.

The Alberta Journal of Educational Research *Vol. XL, No. 2, June 1994, 109-126*

Thomas Maguire
University of Alberta

John Hattie
University of Western Australia

and

Brian Haig
University of Canterbury

Construct Validity and Achievement Assessment

The psychometric notion of construct validity has evolved greatly since its introduction by Cronbach and Meehl (1955). Its current status is to be found in a carefully formulated article by Messick (1989) in which he embeds construct validity in a realist conception of social science. In this article the authors examine several facets of Messick's approach. They take issue with the emphasis placed on consequential validity as a component of construct validity, the inadequate treatment of the link between tests and scores, and the focus on construct validity in large-scale standardized test development. The authors show the profitability of separating the construct-indicator link from the indicator-score link and suggest that although the latter is not always necessary, its treatment in the construct validity literature has generally been inadequate.

Depuis les années 50, lorsque Cronback et Meehl (1955) ont introduit la notion psychométrique de la validité de la synthèse d'éléments constituant une construction intellectuelle, celle-ci a évolué grandement. Cette notion est présentée de nos jours de façon méticuleuse dans l'étude de Messick (1989) dans laquelle il présente cette notion dans le contexte d'un concept réaliste des sciences sociales. Dans cette étude, les auteurs examinent les diverses approches de Messick. Ils questionnent l'emphase que l'on met sur la façon dont la validité conséquentielle ou indirecte peut être une composante de la synthèse d'éléments constituant une construction intellectuelle. Ils questionnent également le traitement inadéquat des liens entre le testing et les scores. Ils interrogent le rôle que joue la validité de la synthèse constituant une construction intellectuelle dans le développement de tests standardisés sur une grande échelle. Les auteurs démontrent à quel point il est profitable de séparer les liens qui déterminent les indicateurs d'une construction intellectuelle des liens indicateurs des scores. Ils suggèrent que l'usage des indicateurs des scores, même s'il n'est pas toujours nécessaire, a tout de même été mal vu dans les écrits qui traitent la validité de la synthèse d'éléments constituant une construction intellectuelle.

Thomas Maguire is a professor of educational psychology. His research interests are in educational measurement and applied statistics.
John Hattie is a professor of education. His research interests are in educational measurement, self-esteem, and multivariate analysis.
Brian Haig is a senior lecturer in psychology. His interests are in philosophy of social science and in the study of research methods.

Messick's (1989) chapter in Linn's *Educational Measurement* must be taken as the most extensive explication of current thinking on construct validity. It aims at providing a "unified though faceted" approach to the notion of validity grounded in ideas about scientific methodology but located in the reality of the political context. The purpose of this article is to examine Messick's notion of validity as it applies to the assessment of educational achievement. Woven through our article are five related points of contention that we have with Messick's chapter as it applies to the construction, interpretation, and use of achievement tests. First, we believe that in spite of the lengthy description of his preferred realist philosophy of construct validity, Messick uses the approach inconsistently. Second, we contend that widening the notion of validity to encompass the notion of consequential validity is inappropriate. Third, Messick's chapter is unsatisfactory with respect to its emphasis on "scores and measurements as opposed to tests or instruments because the properties that signify adequate assessment are properties of scores not tests." Fourth, treatment of structural validity is inadequate. Fifth, the role of the measurement "establishment" is overemphasized at the cost of public accessibility.

The assessment of educational achievement and in particular the assessment of cognition is essentially about describing the unobservable. The focus of such assessment is on internal knowledge states, mental structures, and reasoning processes, and the branch of psychometrics that informs this work is construct validity. We make no apologies for writing this article as a commentary on Messick's (1989) chapter because we believe the chapter is a significant piece and will be a major reference for evaluating the outcomes of research in assessment of cognition.

The article is organized into five sections. In the first, features of Messick's chapter that pertain to our five basic assertions are presented. The second section examines the notion of consequential validity and takes the position that the presuppositions of consequential validity are too important to be treated just as a part of construct validity. A brief discussion of the substantive and external components of validity are presented in the third section, "From Construct to Indicator and Back." Following this, the section "From Indicator to Score" raises issues concerning the underemphasized, overlooked, and misunderstood aspects of structural validity. The concluding section attempts to draw the threads together.

Some Features of Messick's View of Construct Validity

It is fitting to begin with Messick's (1989) opening paragraph because it captures the essentials of both the connotative and denotative aspects of validity for him:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. As will be delineated shortly, the term test score is used generically here in its broadest sense to mean any observed consistency, not just on tests as ordinarily conceived, but on any means of observing or documenting consistent behaviors or attributes. Broadly speaking, then, validity is an inductive summary of both existing evidence for and the potential consequences of score interpretation and use. Hence what is to be validated is not the test or observation device as such

> but the inferences derived from test scores or other indicators—inferences about score meaning or interpretation and about the implications for action that the interpretation entails. (p. 13).

For this article it is important to notice two central features in this quotation: test scores and potential consequences. The main issues of contention derive from the prominence that Messick assigns to each. Although "score" is used in its "most general sense of any coding or summarization of observed consistencies on a test questionnaire, observation procedure or other assessment device" (p. 14), it is important to remind ourselves that the ordinary meaning of the term and its most common application is as a special kind of *inductive summary*, a numerical summary or test score that results from a measurement. The title of the volume, *Educational Measurement*, tells us that we cannot ignore the importance of score in this conventional sense. What readers of previous versions of validity chapters in *Educational Measurement* will find surprising is the importance that Messick assigns to "the functional worth of scores in terms of social consequences of their use" (p. 13).

The unifying structure for Messick's chapter is the two-dimensional Facets of Validity Table in which the "Functions or outcomes of testing" (split into test interpretation and test use), are crossed with the "Sources of justification" (divided into evidential bases and consequential bases) to produce a fourfold table: (a) the evidential basis of test interpretation—construct validity, (b) the evidential basis of test use—construct validity incorporating relevance and utility, (c) the consequential basis of test interpretation—construct validity and value implications, and (d) the consequential basis of test use—construct validity, relevance, utility, and the value and social consequences. Messick talks about the cells as being a kind of progressive matrix moving from the classical Cronbach and Meehl (1955), and Loevinger (1957) approach to validity toward a conception that includes social implications. Although both evidential and consequential aspects of test interpretation precede their appearance as test use in Messick's presentation, we (like Shepard, 1993) prefer the sequential portrayal because it makes it easier to place a boundary between construct validity and what we believe is more properly a matter for professional ethics and social debate.

Test validity has its roots in the positivist-operationalist tradition, where intelligence and personality attributes are characterized in terms of testing operations. In a lengthy section, Messick shows how construct validity, the essence of all validity, is best embedded in a constructive-realist philosophy of science. According to him, "Realists view theoretical terms as conjectures about existing, though presently unobserved, attributes of the world" (1989, p. 26). A construct is the name that we give to the underlying mechanism that explains individual and social behavior, and so for the realist, test and non-test consistencies are explained by reference to a theoretical system of constructs. Construct validation aims at describing, in terms of the surrounding theory, the links between performance on assessment instruments, the nature and construction of the instruments, and the underlying structures and processes that are elicited by the instruments. Although it is possible to talk about validity without reference to a realist position, most current writing in the area is based on that perspective.

The current interest in relating research in cognitive psychology to educational assessment favors a realist position. If we adopt the stance that educational tests should be consistent with the goals and processes of education, if these goals and processes refer to mental and social entities called achievement, and if achievement is presumed to exist and serve an explanatory function (neither a hazy conversational metaphor like "being well dressed," nor a simple operationalist indicator like "number of on-task behaviors") that is in some way measurable, then we are in the realist camp, and the nature of achievement as a construct and its validation warrant our attention.

As it should, Messick's evidential basis of test interpretation borrows much from Cronbach and Meehl (1955) and the subsequent 40 years of construct validity literature. The three components of validity: substantive, structural, and external (Loevinger, 1957); the notion of construct validation as a continuing research process (Cronbach, 1971); the multitrait, multimethod approach to convergent and discriminant evidence (Campbell & Fiske, 1959); and the importance of "domain" (Lennon, 1956) are all to be found in this important section. To the methods of validation listed by Cronbach (1971) Messick adds discourse analyses of various sorts, confirmatory factor analysis, mathematical modeling, and even the analysis of eye movements. But the bulk of the discussion seems to ignore, or at least gloss over, some of the critical aspects: test items seem to appear by magic and are treated as givens; their combination into scores ignores the representational assumptions that would be required by a genuine realist approach; and the emphasis on modeling and components of test variance encourages our attention to drift from achievement about individuals to achievement as an abstraction about groups.

Although the section on the evidential basis of test use covering curricular validity, predictive validity, and utility is a friendly reminder of the psychometrics of Taylor and Russell (1939), Cureton (1951), Cronbach and Gleser (1965), and Ghiselli (1966), the sections on the consequential basis of test interpretation and the consequential basis of test use are a strong plea from Messick to accommodate the various social critics of educational testing by bringing the value orientations, ideologies, social consequences, side effects, and the ethics of testing into the construct validity arena. This is a major point of Messick's work and it is examined in the next section.

Consequential Validity

Messick (1989) claims that consequential validity involves the appraisal of both potential and actual social consequences of test score interpretation and use. It relates to the appraisal of the value implications of the construct label and of the theory underlying the test interpretation, as well as an appraisal of the ideologies in which the theory is embedded. In effect he argues that both the intended consequences and the side effects as seen from multiple perspectives should serve to govern our activities with respect to development, interpretation and use. It is our contention that the priority Messick assigns to consequential validity derives from a particular and narrow political social context, is inconsistent with the realist perspective he describes and endorses, and inappropriately locates certain aspects of professional responsibility with the test developer rather than the test user.

Although following Messick's advice might reduce liability for commercial test developers in a litigious society, such efforts to avoid suits would probably divert energy away from the developers' primary responsibility to collect and report evidence on how the test and its scoring system relate to the underlying construct. In addition, although passing acknowledgment is paid to the test user's responsibility ("a heavy ethical burden,") Messick does not spell this out at all. We believe that Messick's argument for consequential validity as central to validity fails for several reasons.

1. *There is an overemphasis on the role of the developer of large-scale standardized tests.*

If Linn (1989) and his associates had organized their book from the perspective of the test taker rather than developer, it would have been obvious that the CTBS, NAEP, TIMMS, and all the other externally developed achievement measures occupy a relatively small place in the testing lives of students. Teacher-built achievement instruments account for much more of this time, and one would expect the chapter on test validity in one of the significant source books on educational measurement to acknowledge this. The issues Messick raises about consequential validity of standardized tests when applied at the classroom level are really matters for the teaching profession to consider. From the point of view of individual teachers the consequences of their own tests for instruction and student learning are much more important.

Even if we concentrate on standardized instruments, Messick ignores how achievement tests come to be used. In many jurisdictions, national, provincial-state, or local tests are mandated, and they are constructed for specific uses such as accountability. It is true that test developers are often an external agency, but the constraints provided by the test users (i.e., the people or institutions that commission the tests) often set the ideological agenda for test development. So, for example, much of the political support (and funding) for the various international testing programs is to be found in the belief that countries must be competitive in a global economy. Governments think that by knowing how our children compare with the children of other industrial countries in relation to mathematics and science achievement, they can use the information to improve the relative economic position. This forms part of the ideological context that governs the use of scores. The test constructors take great pains in producing instruments that are culturally neutral, useful, and efficient. The values that drive the interpretation do need to be critically examined, but little of this examination will be directed at the activities of the test developers.

2. *There is an emphasis on large-scale, systematic and planned assessments of consequential validity.*

Messick's list of strategies for examining consequential validity makes sense if testing is viewed as an industry like the pharmaceutical industry, but as argued in 1 the bulk of student testing experiences are not of this sort. We believe that the role of test developer (commercial, public, or classroom) is to make the evidential basis for interpretation clear and not to make a list of approved and disapproved uses.

The implicit value assumptions and social consequences of testing are better examined through processes such as those raised by Kane (1992) who presented an argument-based approach to validity that is positioned to deal with how tests should be used and interpreted. Kane suggests that validation is argument that focuses on those parts of the interpretation of a construct that are most doubtful or problematic. He distinguishes two stages. The formative stage involves the clarification or explicit definition of the interpretive argument and the development of a preliminary case for its plausibility, that is, to "develop a plausible conjecture." The summative stage occurs when the interpretive argument is subjected to serious empirical challenges. The degree of confidence in the validity that is warranted depends on how well the interpretive argument survives the various challenges.

To illustrate Kane's (1992) ideas, consider a hypothetical example (although one that resembles an argument current in one provincial department of education). Suppose it is argued that a provincial achievement test in history is presented from a male perspective, such that it has an ideological gender bias. Sharpening this argument leads to two interrelated notions: that boys organize their thoughts on historical matters in a different way from girls, and that the test has been constructed to favor the male "approach." The difference in organization could be due to inherent cognitive differences, differential response to instruction, differences in motivation, differential responses to the test items, and so forth. Examination of the construction process (i.e., the activities engaged in by the test development committee) and statistical analyses of certain items suggest that the criticism is plausible. Using a think-aloud procedure to collect the reasoning processes with carefully selected boys and girls might subject the charge of gender bias to a genuine empirical challenge, and provide some insight into whether curriculum, instruction, and assessment should change, and how that change might be accomplished.

This procedure may not seem to be a reasonable part of the test validation process, because it depends on the test's use in a particular educational-social context to sensitize individuals to the potential problem. Yet once noted, the argument and its evidence could easily change testing practice. Underlying our position is the view that test interpretation is always a matter for debate. The goal of the test development enterprise is to make that debate available to developers, users (consumers), and even test takers where possible, by providing a clear description of the construct and explaining the evidence that links the scores to the construct.

3. *The emphasis on consequential validity favors test score use rather than test development.*

Early in his chapter Messick (1989) notes the importance of scores as "opposed to tests or instruments." Older conceptions of validity spoke of locating tests and constructs in law-like networks (nomological nets). As Shepard (1993) puts it,

> Old conceptions of validity were analogous to truth in labelling standards. A more apt metaphor for current validity requirements is the Federal Drug Administration's standards for research on a new drug. The scientist is responsible for evaluating both the theoretically anticipated effects and the side effects of

> any product before it is rendered "safe and effective" and released for wide-scale use. (p. 426)

This move from test to test use is too great. Even if we accepted consequential validity as an important component of validity, the process of test development would tell us much about the ideological bases of the test and its subsequent interpretation. But Messick's claim that validity is an "integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (1989, p. 13) tells us that what we do with the construct in a utilitarian sense is more important than what it is. In effect, Messick forsakes operationalism for realism only to succumb to a narrow pragmatic instrumentalism. From a realist point of view, a construct denotes an explanatory mechanism, and it may or may not have instrumental application at a particular time. However, our success in understanding the relevant educational phenomena depends in a large part on our effective development of the construct and its associated theory. If there is to be a good test, then an effective test-construct relationship is imperative. For this we must have good constructs. In the present context, achievement as a good construct demands appropriate theory linking knowledge, structures, reasoning processes, and so forth. Tests must derive from such theory, and their use must be consistent with it. However, as we shall see, there are many pitfalls to be negotiated before we get to test scores and their use.

4. *Much of the stress on consequential validity seems to be a reaction to the increased litigation surrounding tests and testing practice.*

Linn (1989) devotes about one quarter of his introduction to *Educational Measurement* to a discussion of the legal aspects of educational measurement, and makes a persuasive case that these concerns are vital to the field of educational measurement. Messick acknowledges this briefly toward the end of his chapter, and although he does not specifically locate his consequential arguments in this climate, it appears to have influenced his decision to assign consequential validity a role of major importance. But this provincialist lapse detracts from the realist position espoused by Messick at the beginning of his chapter. The realist is primarily concerned with fashioning good explanatory theories, and a concern with consequences will relate to that goal, but the greater emphasis will be on coherence, explanatory richness, and so forth.

We would not deny that the consequences of test use are important. Education in general and educational assessment in particular are and should be subject to social and legal scrutiny. But it is our contention that a concern with consequences should be moved out from the umbrella of construct validity and into the arena of informed social debate and formulated into ethical guidelines such as the *Principles of Fair Student Assessment Practices for Education in Canada* (1993). The *Principles* deal not only with items contained under Messick's consequential validity, but speak of follow-up and redress as well. They place ethical test use properly in the arena of professional responsibility and encourage an atmosphere of openness and questioning. If the debate on educational assessments is to be well informed, and if a social science of education is to become more than an oxymoron, we will need to devote our

psychometric attention to the evidential bases for test interpretation and test use. In turn the debate on consequences will rightly guide our subsequent test development activities just as it has with the "authentic" testing movement.

From Construct to Indicator and Back

In this section, we examine the link between constructs and their manifestations. We use the term *indicator* to refer to the responses or behavioral manifestations that might be elicited by an instrument, an observation, or any other testing situation. Indicators can be oral explanations, choice of alternatives in multiple-choice situations, laboratory performances, and so forth. We specifically exclude scores. The link between indicator and score is treated in the next section.

The validity of the link between indicator and construct has largely been studied by examining test performances and arguing back to their validity as signs or samples of the underlying achievement construct. Much significant work has been done in this area, and Messick (1989) acknowledges the importance of cognitive psychologists, the factor analysts, and the mathematical modelers. We are inclined to see the contributions of the first group as being more significant than the others because, as Manicas (1987) notes, causes are not likely to be additive, as the causal modelers commonly suppose. Models like linear factor analysis, or Embretson's (1983) model for response processes assume structural relations that mostly do not exist. Also, we must constantly be mindful that the hypothetico-deductive approach to construct testing cannot carry the entire inferential burden because a single data set can be consistent with two or more contradictory models. Moreover, the emphasis on validation evidence flowing from indicator to construct may subtly encourage an operationalist view of constructs wherein the indicator actually *becomes* the construct. Realists keep the construct/indicator distinction firmly in mind. The construct is the underlying mechanism that explains not only the behavior on a specific indicator, but other, non-test manifestations as well. We argue that the justification process needs to flow in both directions if a coherent realist view of achievement is to be sustained.

There is much to be said about the validity of the link between construct and indicator, but we discuss only two points. The first is the problem of legitimizing the developmental procedures involved in going from construct to indicator. The second concerns the constraints of test format on validity.

From Construct to Indicator

Traditionally, achievement test developers have been advised to adopt a table of specifications strategy that cross-tabulates Bloom, Englehart, Furst, Hill, and Krathwohl (1956) taxonomic categories with content areas. Items are then constructed and placed in categories in proportion to a desired emphasis. This encourages efforts toward content coverage on the part of the test developer, but it fails to take into account some fairly fundamental problems. For example, the Bloom model takes a static view of content, treating it as independent of cognition. As a result, issues such as context salience, position of specific pieces of content in the instructional flow, their position in the overall content structure, and the essential interactive nature of cognition and content tend to be ignored.

More prescriptive approaches to item design that begin with domain specifications and item construction strategies (e.g., Bormuth, 1970; Guttman, 1969; Hively, Patterson, & Page, 1968; Osburn, 1968) are denigrated by Messick (1989) because they "merely tell us at most that the test measures what the universe measures" (p. 40). Yet, on the face of it, it would seem that item generation procedures should speak to the validity of the link between construct and indicator.

Haertel (1985) was one of the first people to examine achievement as a construct. He argues that educational outcomes should be thought of as constructs:

> They are artificial, designed deliberately to equip students with a repertoire of appropriate responses to the complex settings and symbols of our culture. They are attributes defined primarily in terms of their behavioral manifestations and only secondarily in terms of cognitive processes or memory mechanisms. (p. 28)

Despite appearances this does not satisfy a realist orientation. Haertel's second sentence gives the game away. From a realist perspective an educational outcome would not be seen as an explanatory construct. Achievement as a construct would be described *primarily* in terms of internal processes and mechanisms, and these would be seen to give rise to certain behavioral manifestations. Educational *outcomes* are obviously important, and they are often described in behavioral terms. But the cardinal goals of education are tied to the development of internal achievement constructs. Snow (1989), for example, lists five broad achievement constructs: deep understanding, efficient intuitive use, multiple flexible strategies, adaptive action control, and achievement motivation.

For the realist the main problem for assessment is to develop and link indicators to the networks of such constructs. One approach to achievement assessment that exemplifies something of the realist view is Biggs and Collis's (1982) SOLO taxonomy. SOLO, the Structure Of Learning Outcomes, is a system for categorizing student responses according to their degree of elaboration, organization, consistency, capacity, generalization, and abstraction. The levels of the taxonomy can be thought of as stages of learning that an individual passes through in mastering new concepts. The levels of the taxonomy used to evaluate the quality of learning are described in terms of the amount of memory capacity required to make the response, the kind of relating operations used to produce the response, and the degree of consistency and the quality of closure displayed in the response. The application of the SOLO hierarchy is centered on declarative and procedural knowledge, but the attainment of higher levels presumes appropriate motivational orientations, self-regulatory functions, and learning strategies. The taxonomy is one promising attempt to represent the relationship between knowledge and its organization as the student is prepared to present it under test conditions. Moreover, it is a framework that can be used from classroom tests to national assessments.

The Constraints of Test Format

In the previous section, we talk about the construct-indicator link. In the case of achievement measurement, the indicator is usually a response to a task or an item. Like many others, we believe that test format can influence a test taker's

approach. The format influence may be a source of invalidity, but it is not self-evident. If we place formats along a continuum according to the degree to which the test taker's response is apparently constrained, multiple-choice questions would fall at one end and performance tasks would fall toward the other. But constraint on response may be inversely related to an item's "representational validity" (how well it represents the achievement construct). Complex performance-based assessments place great emphasis on problem solving, critical thinking in context, and other "higher order" thinking skills. As Linn, Baker, and Dunbar (1991) point out, this may not guarantee validity but it is a good start. Frederiksen and Collins (1989) argue that there are validity advantages to tests that evaluate cognitive skills "directly." Such tests obtain "systemic" validity because instruction that improves test scores will also improve performance on the extended task.

But more constrained formats like multiple-choice forms do not make the link simple. The argument that multiple-choice items only assess recognition simply cannot be credited by anyone who has made the effort to find out what processes students use when responding. To show this clearly, Skakun (1993, personal communication) asked senior medical students to think aloud as they responded to several multiple-choice questions in internal medicine. Three example answers to one of the questions are shown in Figure 1. Reading through the responses it is easy to see that a single question has tapped into an incredible variation in array of internal structures. We believe this illustrates the main problem with the construct validity of the multiple-choice format. Far from being a direct method for assessing recognition, this format is well capable of eliciting internal responses that are extremely complex. The problem is that these responses are not entirely foreseeable. The positive features of multiple-choice items are well known. But the one that is most supportable may be the efficiency of content coverage. If this feature is highly valued in a specific context, Willson's (1989) advice is to link each of the distractors to the process and knowledge requirements of the achievement construct.

As Snow and Lohman (1989) tell us, much important work is being done on the validity of the construct-indicator link. Our caution is to keep the bidirectionality of evidence in mind when thinking about this link, and to be aware of how the form of the indicator might influence the way it links with the underlying construct.

From Indicator to Score

One of Loevinger's (1957) three components of validity is structural validity. She describes it as the

> extent to which structural relations between test items parallel the structural relations of other manifestations of the trait being measured.... [it] includes both the fidelity of the structural model to the structural characteristics of non-test manifestations of the trait and the degree of inter-item structure. (p. 98)

This means that internal characterization of the construct should be represented in the indicator. For example, if we accept the metaphorical view of achievement as milestones along an educational highway, then the achievement indicators should form a scale. But if we see achievement as the emergence of stages through which an individual passes as he or she masters,

Item

A 56-year-old man presents to his doctor with a month history of intermittent facial pain. On examination he is found to have a diminished corneal reflex and slight hearing defect on the right. The diagnosis is:

1. Right cerebral tumor
2. Trigeminal neuralgia
3. Otitis media
4. *Acoustic neuroma
5. Multiple sclerosis

Response from Examinee A:

56-year-old man, they want a diagnosis, comes to doctor one month, history of right facial pain. I am thinking neurology. Diminished corneal reflex, slight defect in hearing. So we have facial pain on the right and pain is carried by the 5th cranial nerve, diminished corneal reflex again the 5th cranial nerve and he has hearing defect. Hearing defect, it could be sensory or conduction, so it could be related or not. If it is related it would indicate 5th cranial ... I mean 8th cranial nerve so there would be more that I would want to do on examination. I would want to find out how the 7th nerve was—whether he had any sort of ocular—extraocular defects—motor component but I guess if all they found on examination was no extraocular involvement just cranial nerve 5, I would go to the choices right now. Right facial pain—Right cerebral tumor, trigeminal neuralgia, otitis media, acoustic neuroma, multiple sclerosis. His age is getting up there for MS, acoustic neuroma has nothing to do with this corneal reflex unless it's impinging on the 5th cranial nerve ... one month history of intermittent facial pain, I'm going back up because obviously I don't know the answer right off, so I'm looking at the stem for clues I missed—like intermittent and diminished corneal reflex and slight defect on hearing so these are all, they are not cut and dried symptoms, they come and go or just decrease so it could be conduction problems and they are separated. I would want to know if he has had a past history of anything like this. Trigeminal neuralgia ... well, I know tic douleureux can present with facial pain but I haven't heard of hearing defect with it, but it's on the differential. So right now the cerebral tumor ... no, because it's in the head, up top, supratentorial and I'm looking a a cranial nerve defect. Trigeminal neuralgia—that's possible. It only usually ... it's usually the mandibular branch of the 5th cranial nerve. Otitis media, no there wouldn't be a problem with the corneal reflex. Acoustic neuroma, it's possible, but it shouldn't affect the 5th cranial nerve. Multiple sclerosis, that's on my differential so I'm stuck between trigeminal neuralgia and multiple sclerosis. I'll pick multiple sclerosis because of the distance between the nerves affected. I don't see the 5th cranial nerve coming out of the pons traversing the skull going through the foramina—I don't see it associated.

Response from Examinee B

A 56-year-old man, so he's middle-aged. He has a one month history of intermittent facial pain, to me ... right away I think of a trigeminal ... some kind of neuralgia. On examination he's found to have a diminished corneal reflex, so to me that means his 5th cranial nerve is involved. And slight hearing defect on the right and to me hearing defect in an 8th cranial nerve abnormality, and facial pain is also 5th nerve. So his 5th nerve and his 8th nerve are involved. So I'm looking in the answers for something that would cause that.

I'm looking at the answers and number 5, I don't think it's multiple sclerosis, its not typical presentation of multiple sclerosis. Otitis media, I'm looking at number 3, I don't think that's likely, the symptoms don't really correspond with that. Trigeminal neuralgia, I think it could be except for the fact that he's got hearing involved. So, and right cerebral tumor, it seems like it's cranial nerves that are involved and not the cerebrum so I would exclude that. So I would say number 4, acoustic neuroma, but I don't know why his facial nerve isn't affected, nerve number 7, but I would say either number 2 or number 4 and I'm going to pick number 4 because number 2 involves only the 5th nerve. Then again, I can't decide because it looks like ... I'm having a hard time deciding between answer number 2 and answer number 4 cause it looks like it could be ... like it could be trigeminal neuralgia except that he's got a hearing defect but that could be a

Figure 1. Example item and think-aloud responses.

separate thing. So, but I think neuralgia only leads to pain and not really diminished corneal reflex, I'm not sure but I'm going to stick with number 4 still.

Response from Examinee C

A middle-aged man with right facial pain and on examination diminished corneal reflex and slight hearing defect ion the right. Um, anytime there is an asymmetrical ... here in the absence of ... acoustic neuroma which is down here in number 4.

Right cerebral tumor is not in the right area to be cerebellar. Trigeminal neuralgia, there should be pain—there is some pain but it wouldn't explain the hearing defect. Multiple sclerosis is possible but unlikely in a gentleman of this age. Otitis media—unlikely. Acoustic neuroma in this gentleman until proven otherwise—number 4.

Response from Examinee D

A 56-year-old man presents ... slight hearing defect on the right. So it doesn't say where the right facial pain is caused if it's trigeminal neuralgia it would be greater in V2 than V3 then V1. He has diminished corneal reflex and a slight hearing defect on the right—hearing defect so 5 and cranial nerve 8 are being affected. Diagnosis of right cerebral tumor, no I don't think you can say that just from the history, number 1 is out, number 2 is possible, number 3, otitis media, well it doesn't explain the neurological findings, number 4 acoustic neuroma, boy oh boy, that's a possibility, number 5 MS he's not in the right age group. So between 2 and number 5—sorry 2 and 4, and a slight hearing defect on the right, I think with acoustic neuroma that hearing defect would probably come on before the corneal reflex, I think it would be pretty marked if you had a diminished corneal reflex so I'll say number 2.

Figure 1 (continued).

organizes, elaborates, and abstracts declarative and procedural knowledge, then the indicators should form a some kind of linked hierarchy. That Messick's (1989) comments on the indicator-score relationship are essentially those of Loevinger gives the misleading impression that this aspect of construct validity is not problematic. This is unfortunate because it is the central issue in the *measurement* of achievement (i.e., the link between indicator and score). Much effort has been expended on developing a technology of educational measurement, as witnessed by the large portion of journal space devoted to item response theory and related topics, but surprisingly little attempt has been made to integrate that technology with the overriding concerns of construct validity. Before expanding on this point, it is useful to step back and raise some questions about basic measurement theory.

Let us take as an initial characterization of measurement, the mapping of objects or phenomena onto a numerical continuum according to the variations in their characteristics. There is much we can do when we have described phenomena in numerical terms, and the rewards are so compelling that we often neglect to ask the important prior question, Is this phenomenon measurable? To answer the question, we need to know much about the object or phenomena, and we need to have a position on the nature of the mapping rules. As Michell (1990) points out, there are two conceptions of measurement: operationalism and representationalism. According to operationalism, concepts are indistinguishable from their corresponding sets of operations; measurement is simply a set of procedures that generates a number. From a realist point of view, this is not an acceptable position.

According to Michell (1990), under the representational theory of measurement, "Numbers are assigned to empirical entities in such a way that certain relations between the numbers assigned represent empirical relations between

the entities. Representationalists make a distinction between empirical entities and numerical ones" (p. 29). As applied to achievement, this means that there is a distinction between a test score and its construct. Berka (1983) makes the obvious distinction between physical measurement and extraphysical measurement as is found in the social and behavioral sciences. Measurement of constructs like achievement clearly fall in the latter category. Berka points out that extraphysical measurement is methodologically tied up with classification (dividing objects into sets according to some qualitative aspects) and with observation, and although it proceeds in a different manner from the usual experimental basis of physical measure, there are three common components: the object of measurement (in our case achievement), the results of the measurement (the value on the scale), and the instrument or the mediating empirical operations (for us the achievement test and its scoring system). If we are to work from Berka, we might be advised to keep the prior notion of classification in mind as we move toward valid "mediating empirical operations" (i.e., tests!).

Scoring models from classical test theory and from item response theory (IRT) provide two popular but distinct structural models of achievement. Under classical test theory, building tests from discrete items and adding the item scores together is consistent with a metaphorical view of achievement as a long coin tube, where each piece of achievement is a coin. A student's score is given by the number of achievement coins that he or she possesses. The presence of this or that coin and their order in the tube are not important; it is the sum of the item scores that indicates the amount of achievement. The IRT version of achievement is like the highway described earlier. Items are like milestones along the road. The achievement score tells us at which position the student is to be found. Under the Rasch scoring rules of IRT, these summative and positional approaches yield the same results, but that is a consequence of the mathematical assumptions of the procedure and not an ontological assumption about achievement itself.

To appreciate the distinction, consider the example of the Alberta Infant Motor Development Scale (Piper, Pinnell, Darrah, Maguire, & Byrne, 1992). The scale is designed to assess motor development in infants from birth to independent walking. The items form an almost perfect Guttman scale. Interestingly enough, however, if an infant is located at a particular point on the scale he or she may no longer be able to successfully perform items from a previous point. Those items (i.e., earlier items) may have been fundamental to healthy development, but they are lost or incorporated into higher level skills. Total scores on items would yield different positions for infants than milestone scores. We know that the latter has greater construct validity because it conforms to the most coherent theories of infant motor emergence.

Examples of achievement constructs that fit either the classical summative scoring model or the item response continuum model are less common than the advocates of either approach would have us believe. Without surrendering to the operationalist camp, we agree that observables and the empirical models on which they are based must influence the notion of the construct, but in the case of most testing programs this has gone too far. The scoring model en-

courages a view of achievement that is simply inconsistent with most of our current understanding of achievement.

Portrayals of the achievement construct such as that given by Snow (1989) prompts us to question whether achievement is measurable at all. Remembering Berka, we should think in terms of classification of achievement types. Perhaps the criterion-reference people have been correct all along with their decision approach to achievement assessment. Although it did not take long to convert the simple ideas of Glaser (1963) into the mathematics of cut-off scores, loss functions, and domain estimates (Hambleton, Swaminathan, Algina, & Coulson, 1978), the essential notion of competence is still of use. The configuration, pattern, or position that a person possesses in relation to the achievement construct of interest is characterized via the indicators. Our initial role as assessors is to sort these portrayals into groups that are similar. Judgments can then be made on whether a particular group possesses "mortal deficiencies." This is not yet measurement, but neither does it distort the construct representation due to an inappropriate and arbitrary scoring system.

If we want a measurement, and we suspect that this should be far less often than current practice suggests, we need to think about defining a single continuum. Achievement as Snow (1989) and others construe it is complex and multifaceted, so that indicators of achievements might be ordered differently in relation to different facets. For example, if we wanted to look at achievement with respect to mathematical problem solving and we administered several open-ended mathematical problems, then we might consider facets like efficiency, creativity, transferability, conformity to taught patterns, and so forth as being important, and the performances on these facets would have different positions relative to each other. The scaling solutions would require assumptions to be made, but care could be taken to ensure that they retained a reasonably plausible representation of the facet of the construct.

If decisions require a single ordering of people, then the nature of the decision should guide the scaling process. If the decision requires only categories, then the combining of indicators need not produce a score. Take a complex example to illustrate. Each year the Medical Council of Canada administers a licensing examination to about 1,700 graduating medical students from across the country. In addition, approximately 700 graduates from outside Canada take the exam. Three quarters of the examination is made up of multiple-choice questions assessing competence in six disciplines (medicine, surgery, pediatrics, obstetrics and gynecology, psychiatry, and preventive medicine and community health); the remainder of the examination consists of patient management problems. Until recently, for candidates who wrote the exam, the number correct on the multiple-choice section was combined with appropriate weighting to the patient management section, and the candidates were ordered. The 4th percentile of the Canadian student performance was used as the licensing criterion.

If we think of the multiple-choice portion of the examination as a large achievement test, with each item testing some complex aspect of medical training, we might ask ourselves if there is a more valid way of arriving at the decision to license or not. Notice that there is no inherent reason for recasting the problem as a cutting score problem.

The multiple-choice items are selected from a bank or are created by the six disciplinary test committees. Each committee is composed of six members who are a careful blend of discipline specialists, specialists in family medicine, faculty of medical schools, and general practitioners. Each of these committees participated in a modification of the Nedelsky standard setting procedure using a kind of Delphi approach (Maguire, Skakun, & Harley, 1992). The goal of the process was to reach consensus judgment on each alternative to each item. The basis of the judgment was that a graduating medical student who is about to enter supervised clinical practice should know whether an alternative is plausible or not. Without apology, the arguments and judgments were based on the experience of the panels, the importance of the concept (e.g., is it fundamental, is it dangerous to be wrong about it?), and the knowledge that passing candidates would be allowed to take internships anywhere in the country.

As noted above, each of the items is capable of eliciting a complex sequence of thought and there are many ways of getting an answer correct. But if the thrust of the decision is to eliminate people with inappropriate conceptions, a scoring system could have been built up from patterns of responses on these items, so that, for example, it might turn out that the panels believe that any candidate who makes a particular pattern of errors should not be granted a license to practice. Several patterns of responses might be judged inadequate, and adjustments could be made to fit the uncertainties of the testing situation. If it turned out that a cutting score on the simple number correct yielded the same decisions (and it might), then much effort could be saved by doing that. But the justification would lie in the fact that it achieved the same results as a process that retains representational fidelity.

Essentially our thesis is that tests and scoring procedures are *mediating empirical operations* to make constructs into numbers. If the things that we do with those numbers (make decisions, conduct research, etc.) are to have any validity, then we must take great care in developing the tests and scoring procedures. It is not obvious that the common practice of administering large numbers of items and computing number correct, or weighted aggregates, retains the representational link between scores and constructs. Given the complexity of achievement constructs we think it will be often more useful to think in terms of ordered categories.

Concluding Thoughts on the Construct Validity of Achievement Tests

At the end of the day we must ask ourselves if the realist view of achievement is worth the effort. We think the answer is Yes. This is because of our belief that educational social science advances on the strength of well-developed theoretical constructs. The importance of achievement stems from it being more than an indicator; it essentially involves changes in the way students think and understand. Pursuing achievement indicators that reflect deeper structural changes, that is, that possess construct validity, is worthy of our attention.

We have tried to establish the importance of two links: construct-indicator, and indicator-score. Although they are blurred in most of the educational measurement literature, we think the distinction between them is useful in helping us to see the greater importance of the first. In considering these links we are led to the conclusion that many scoring systems in use have not been

validated in Loevinger's (1957) structural sense, and that it may not be necessary. However the urge to move to scores is ingrained in our thinking. The test validation procedures in common use assume a score. Factor analysis, the traditional workhorse of test validation, and structural modeling, its likely successor, seem like powerful procedures. But they have their limitations. They make metric assumptions that have consequences for our understanding of the nature of the relationships among the constructs they help us to investigate, but these assumptions are wantonly disregarded in practice. In addition, they tend to encourage sloppy talk about the nature of causal relationships.

The center of that sloppiness is to be found in the notion of variance. In the Messick (1989) chapter its use is pervasive. Shared variance, test-irrelevant variance, and so forth are central to the discussion of evidential aspects of validity. But to be pedantic for a moment, variance is a description of dispersion in a group of scores. Even at a superficial level, to talk about sharing variance, or apportioning variance, requires adherence to a particular kind of model (usually additive) that may well misleadingly depict how constructs relate to each other in an individual or a group of similar people.

We believe that the enlightened study of achievement and how it evolves will involve the intensive study of one or a few people at a time. Alton-Lee and Nuthall (1992) conducted a series of studies that provide an excellent illustration. They observed and recorded all the student teacher interactions, student-student interactions, and individual student vocalizations and sub-vocalizations that occurred during the teaching of an entire unit of instruction. They used four different units of instruction (ranging in time from 5 days to 31 days) and reported their results on 14 different students (3 or 4 per unit). By looking at pre-test, immediate posttest and 12-month posttest changes on achievement items (from 81 to 113 per unit), they classified outcomes as: items that were already known; items that were not learned; items that were learned and forgotten; and items that were learned and remembered. From their observations on two students, they produced a model that described student learning as the creation of specific knowledge constructs. The model defined the number of instances and types of relevant experience required for construct generation and the time intervals in which the experiences must occur. Using the model on other students they were able to predict with over 80% success student learning (or not) on individual test items.

Studies like those conducted by Alton-Lee and Nuthall (1992) tell us much about how classroom activity influences learning, but to improve our understanding of achievement we will need to become even more penetrating at the micro level. The strategies of investigation will be as Berka (1983) notes: logical, observational, and qualitative. The gains from this should be worth the effort.

Tests are often thought of as the tail that wags the educational dog. But this metaphor is too glib. The consequences of educational assessment both for individuals and for society are important. Testing works in support of education, and our position underlines the need for their principled use, but construct validation as it applies to achievement assessment should mainly be directed at promoting a real understanding of the nature of achievement constructs and how they might be represented in assessment. To focus on the

consequential use of tests is to encourage a view that educational stakeholders are relieved of their responsibility to scrutinize all facets of the educational enterprise in the context of emerging social change.

References

Alton-Lee, A., & Nuthall, G. (1992, April). *Students learning in classrooms: Curricular, instructional and sociocultural processes influencing student interaction with curriculum content*. Paper presented at the Annual meeting of the American Educational Research Association, San Francisco.

Berka, K. (1983). *Measurement, its concepts, theories and problems*. Boston: Reidel.

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning, The SOLO Taxonomy*. New York: Academic Press.

Bloom, B.S. (Ed.)., Englehart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *A taxonomy of educational objectives: Handbook I: Cognitive Domain*. New York: Longmans, Green.

Bormuth, J.R. (1970). *On the theory of achievement test items*. Chicago: University of Chicago Press.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L.J., & Gleser, G.C. (1965). *Psychological tests and personnel decisions*. Urbana, IL: University of Illinois Press.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cureton, E.E. (1951). Validity. In E.F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). Washington, DC: American Council on Education.

Embretson, W.S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179-197.

Frederiksen, J.R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Ghiselli, E.E. (1966). *The validity of occupational aptitude tests*. New York: Wiley.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18*, 519-521.

Guttman, L. (1969). (Ed.). Integration of test design and analysis. *Proceedings of the 1969 Invitational Conference on Testing Problems* (pp. 53-65). Princeton NJ: Educational Testing Service.

Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research, 55*, 23-46.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D.B. (1978). Criterion-referenced measurement: A review of technical issues and developments. *Review of Educational Research, 48*, 1-48.

Hively, W., Patterson, H.L., & Page, S.H. (1968). A "universe defined " system of arithmetic achievement tests. *Journal of Educational Measurement, 5*, 275-290.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Lennon, R.T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16*, 294-304.

Linn, R.L. (1989). Current perspectives and future directions. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 1-10). New York: American Council on Education and Macmillan.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-23.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694 (Monograph Supp. 9).

Maguire, T.O., Skakun, E.N., & Harley, C. (1992). Setting standards for multiple-choice items in clinical reasoning. *Evaluation and the Health Professions, 15*, 434-452.

Manicas, P.T. (1987). *A history and philosophy of the social sciences*. Oxford: Blackwell.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Erlbaum.

Osburn, H.G. (1968). Item sampling for achievement testing. *Educational and Psychological Measurement, 28*, 95-104.

Piper, M.C., Pinnell, L.E., Darrah, J., Maguire, T.O., & Byrne, P.J. (1992). Construction and validation of the Alberta Infant Motor Scale. *Canadian Journal of Public Health, 83* (Supp. 2), s46-s50.

Principles of fair student assessment practices for education in Canada. (1993). Edmonton, AB: Joint Advisory Committee. Centre for Research in Applied Measurement and Evaluation, University of Alberta.

Shepard, L.A. (1993) Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education, 19,* 405-450.

Snow, R.E. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher, 18*(9), 8-14.

Snow, R.E., & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Taylor, H.C., & Russell, J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology, 23,* 565-578.

Willson, V.L. (1989). Cognitive and developmental effects on item performance in intelligence and achievement tests for young children. *Journal of Educational Measurement, 26,* 103-119.

The Alberta Journal of Educational Research *Vol. XL, No. 2, June 1994, 127-146*

Gay L. Bisanz

and

Jeffrey Bisanz
University of Alberta

Studying the Development of Academic Skills in the 1990s: Implications for Assessment

Psychological research on the development of academic skills has implications for educational assessment. We review research in two areas that typify the enterprise, arithmetic, and early reading, and we identify some of these implications. One conclusion is that productive interactions between assessment and cognitive development may well depend on the degree to which these specialities collaborate to produce individuals with the skills and insights that can arise from an interest and expertise in both domains.

La recherche psychologique sur le développement des habiletés académiques a une certaine portée sur l'évaluation en éducation. Nous passons en revue deux domaines de la recherche qui traitent de l'évaluation en arithmétique et en lecture hâtive et nous identifions quelques-unes de ces portées. Une conclusion établit que les interactions productives entre l'évaluation et le développement cognitif peuvent bien dépendre du niveau par lequel ces deux spécialités collaborent afin de développer des individus possédant des habiletés et une perspicacité qui proviennent d'un intérêt et d'une expertise dans les deux domaines.

Educational assessment in Canada and the United States is sustained by a variety of stakeholders—including politicians, administrators, teachers, parents, and students—with diverse and sometimes conflicting needs. Some stakeholders are concerned about systemic evaluation for purposes of public accountability. Their goal is to meet public standards for education while minimizing waste and expense. Others are opportunists, using test scores to sell homes in desirable neighborhoods or hawk the latest textbook series. Still others, like parents and teachers, have in mind particular children in their care. They value assessments that will help these children learn in school. Public debate among stakeholders about assessment often seems related to the unlikely but efficient premise that one type of assessment can be developed that will be suitable for all needs. We are skeptical about this premise. Rapid advances in new methods of assessment may well depend on the extent to which stakeholders come to recognize the diversity of needs and support the development of multiple methods of assessment to serve them.

Gay Bisanz is an associate professor of psychology and Director of the Centre for Research in Child Development. She studies cognitive development with special interests in early literacy skills, text comprehension, and science instruction.
Jeff Bisanz is an associate professor of psychology and a fellow of the Centre for Research in Child Development. He studies cognitive development and is engaged in research on the development of mathematical skills and on scientific reasoning.

Our goal is to review recent research on cognitive development with a focus on two areas of study, arithmetic and early reading, and to explore implications of this research for the development of new assessment techniques. Most of this work is focused on processes of knowledge acquisition, including the mental processes involved in cognition and motivation as well as the social and cultural processes that characterize the environments in which children learn. As we describe this research, it becomes clear that the contributions of research on cognitive development to educational assessment are likely to be greatest for stakeholders whose primary concern is the optimization of learning in individual children. These stakeholders include teachers, parents, and the students themselves.

Since the turn of the century and the work of G. Stanley Hall, developmental psychologists have shared an interest with many members of the general public in understanding and optimizing children's growth and development (Cairns, 1983). Widespread work related to the development and optimization of academic skill is relatively recent, however. A review of research on cognitive development in the 1960s would reveal a substantial number of articles on learning using traditional laboratory tasks, as well as a flurry of investigations focused on exploring and extending Piaget's theory. These lines of research were supplemented by the growth of a significant body of work on the development of memory, inspired in part by models of memory in adult cognition but with orientations and methods that reflected the special challenges of studying development rather than performance (e.g., Kail & Hagen, 1977). The use of information processing frameworks and methods was in its infancy, having been suggested by Simon in 1962. Despite growing interest in learning and cognitive development in children, research specifically designed to investigate the acquisition of academic skills was rare. For example, very few pages in Stevenson's (1972) capstone book on children's learning were devoted to research on academic skills.

Today neither traditional learning theory nor Piagetian theory provides a unifying framework for research on cognitive development. Information processing approaches and various contextual theories are important but not dominant. Research on cognitive development in specific domains (e.g., language, mathematical reasoning, naive theories in science) is clearly ascendant, but it is eclectic with respect to theoretical perspective. Despite the focus on development in specific domains, many developmental researchers have not abandoned Piaget's search for unifying principles in development. This agenda is represented, for example, by Kail's (1991) research on age-related changes in processing speed, Case's (1993) search for central conceptual structures, and Siegler's (1986) analyses of common developmental processes across different domains of cognitive skill. Renewed interest in Vygotsky's work and, more generally, on the role of social processes on cognitive development (Wertsch & Kanner, 1992) reflects a growing concern with the imbalance arising from the Piagetian metaphor of child-as-scientist. Establishment of the *International Journal of Behavioral Development* and the growth of the International Society for the Study of Behavioral Development reflect an increased awareness of the importance of culture in the study of cognitive development.

In this setting, developmental psychologists have begun (a) to characterize the acquisition of specific academic skills, such as reading, writing, problem solving, and reasoning in subject matter domains, and (b) to identify the effects of schooling as a set of experiences that shape cognition. The work on arithmetic and early reading typify the enterprise. We review research on arithmetic to illustrate the interplay between methods and theories used in research to assess cognitive skills (see also Royer, Cisero, & Carlo, 1993). Research on early reading is used to illustrate the perspectives afforded by new developmental theories and the factors that affect learning in children's environments. Finally, we examine a number of implications for assessment that arise from this research.

Research on the Development of Arithmetic Skills

Research on the development of mathematical cognition has mushroomed over the past two decades, largely because of methodological and theoretical advances that have enabled investigators to address important questions about the relations among general remembering and problem solving processes and knowledge of material in a specific domain. For purposes of illustration, we focus primarily on research designed to explore how children and adults solve single-digit arithmetic problems. Proficiency in simple arithmetic is only a component of mathematical cognition, but it is common, practical activity that is likely to be important for fluency and motivation in more complex forms of problem solving and reasoning.

Doing Arithmetic

Identifying procedures. The traditional approach to assessing proficiency in arithmetic is heavily oriented to the *products* of children's thought, as opposed to the *processes* that generate products or answers. Children or adults typically are presented with a set of arithmetic problems and the number of problems solved correctly is determined, often within some limited period of time. So, for example, a student who provides the set of answers given in Figure 1 would be judged to have solved 60% of the items correctly.

In contrast to traditional assessments of achievement, cognitive approaches are heavily oriented to identifying the mental processes that underlie performance. The distinction between product and process orientations is highlighted by a careful examination of the answers given to the problems in Figure 1. Notice that all the answers, correct and incorrect, reflect the operation of a consistent, underlying rule, according to which it is appropriate to subtract the smaller value from the larger value, regardless of which number is on top and

(a)	(b)	(c)	(d)	(e)
823	632	259	389	426
-332	-220	-127	-229	-349
511*	412	132	160	123*

Figure 1. Five subtraction problems. Incorrect answers are indicated with an asterisk.

which is below. Erroneous computational rules of this type are often referred to as *bugs* because, like bugs in computer programs, they are responsible for inappropriate output. A number of such subtraction bugs have been identified in children's arithmetic (Brown & Burton, 1978; Brown & VanLehn, 1982). Notice that the extent to which a bug produces erroneous output depends on the problem to which it is applied. Despite the bug, a child might obtain a high score if the problems on a test are similar to problems (b)-(d) in Figure 1. Even if performance is not perfect, a high score might serve to strengthen the bug, especially if the child views arithmetic as a maze of arbitrary rules that can never be negotiated with complete success.

The distinction between product and process is not unique to contemporary cognitive psychology (e.g., Werner, 1937), but cognitive and developmental psychologists have developed a variety of methods to identify the processes used on academic tasks (Kail & Bisanz, 1992). Analyses of response accuracy, similar to our example with subtraction, are instances of a method known as *rule assessment* (Siegler, 1976, 1981). Researchers also examine the solution time or *latency* to identify the procedures used by students. For example, young children often use counting-based procedures to solve simple arithmetic problems. According to a sum procedure, when a child is presented a numerical problem of the form a+b, he or she increments a mental counter a total of a+b times. Each such increment presumably takes a certain amount of time, and so solution latencies should vary as a function of the sum of a and b. According to an alternative hypothesis, the child sets the mental counter to the value of the larger number, a or b, and then increments the mental counter a number of times corresponding to the smaller of the two digits. Consequently, latencies should vary as a function of *min* (a,b), rather than *sum* (a,b). Hypotheses such as these can be tested by comparing latencies for different problems (Ashcraft, 1982).

Finally, analyses of behavioral observations and self-reports are also conducted to identify the procedures used to solve problems. While solving addition problems, for example, children often can be observed to count on their fingers and, when asked, can describe a counting procedure. Self-reports are particularly useful for obtaining insights about solution procedures when children use methods, such as retrieving answers from memory, that are not likely to have behavioral manifestations (e.g., Bisanz, Morrison, & Dunn, in press; Siegler, 1987).

Thus researchers have relied on several different kinds of dependent measures to identify how children solve arithmetic problems. Each of these methods has limits. For example, rule assessment methods can be ineffective if students do not use a single rule consistently (Kail & Bisanz, 1982), analyses of latencies can be qualified by patterns of error data (Pachella, 1974) or by variability in the procedures used (Siegler, 1987), and the validity of self-report data may be limited under conditions that are not as yet well specified (Cooney & Ladd, 1992; Russo, Johnson, & Stephens, 1989). Given these problems, the most compelling conclusions typically arise from studies in which the results from different dependent measures converge (e.g., Siegler & Shrager, 1984).

Three clear conclusions have arisen from this research. First, solution procedures change considerably as children become more skilled. For example,

younger, less skilled children tend to use counting-based procedures, such as the *sum* and *min* procedures, whereas older children and adults tend to rely primarily on direct retrieval of arithmetic facts from memory (e.g., Ashcraft, 1982; Ashcraft & Fierman, 1982). Second, individuals typically use a variety of different solution procedures to solve problems rather than being slavishly attached to a single procedure (Siegler, 1987; Siegler & Shrager, 1984). Even adults sometimes use counting-based procedures rather than retrieval to solve simple addition problems (Geary & Wiley, 1991; LeFevre, Sadesky, & Bisanz, 1994; Thibodeau, 1993), and very young children occasionally use retrieval rather than counting (Siegler & Robinson, 1982). Third, many procedures are developed without direct instruction as children solve problems in and out of school (Bisanz et al., in press; Ginsburg, 1986).

Recognizing that individuals have a repertoire of procedures that can be used adaptively has several important implications for developmental theory and for methods of assessment (Bisanz & LeFevre, 1990; Siegler & Shrager, 1984), two of which are of immediate relevance. One implication is that a level or stage of development is not necessarily defined most accurately in terms of the degree to which a single, dominant procedure is used. In mental arithmetic, it would be erroneous to assume that children at developmental level p use procedure x exclusively, that children at level $p+1$ use procedure y, and so on. The notion that stages or phases of development are defined in terms of single strategies, operations, or concepts has dominated developmental theory for a long time, but clearly this approach can be misleading.

Recognizing that children use diverse procedures to solve problems also implies that if we are to understand developmental change and sources of individual differences in performance, then we need to understand the process by which these procedures are selected. Considerable progress has been made on this issue over the last decade (Ashcraft, 1987; Siegler, 1987, 1988; Siegler & Shipley, in press; Siegler & Shrager, 1984), and a consensual view is emerging that appears to have broad applicability. Here we highlight some of the main characteristics of this emerging view by briefly describing elements of Siegler's (1988; Siegler & Shrager, 1984) theory of the selection process.

Selecting procedures. In Siegler's view, solution of arithmetic problems is a two-phase process (see Figure 2). When presented with a problem, a person first attempts to retrieve an answer directly from his or her memory for arithmetic facts. If that attempt fails, a second phase is engaged in which a backup procedure, such as *min* or *sum*, is used to generate or construct an answer. Thus whether a person uses retrieval or a backup procedure depends on the success of the retrieval phase, which in turn depends largely on knowledge of arithmetic facts.

In Siegler's model, knowledge of arithmetic facts is represented in terms of associative strengths between specific problems (e.g., 4+5) and possible answers (e.g., 6, 7, 8, 9, etc.). Retrieval of an answer from memory is probabilistic and depends on associative strength or activation levels [$Act(A_N)$], so that weakly associated answers to a problem are retrieved infrequently and strongly associated answers are retrieved often. Thus knowledge of arithmetic facts is represented in terms of *distributions of associative strengths* for different problem-answer combinations. Two such hypothetical distributions are illustrated

in Figure 3. Note that for 5+1, the association with one answer (6) is very strong and associations with other answers are weak, whereas for 4+5 none of the candidate answers is associated strongly. Such distributions are characterized by their degree of *peakedness*: the distribution of associative strengths for 5+1 is *peaked*, whereas the distribution for 4+5 is *flat*.

Distributions of associative strengths have an important influence on the selection process. Note in Figure 2 that an individual first encodes the problem and then sets a search parameter (n) that determines how persistent he or she will be in attempting to retrieve a solution. On the first attempt, if the retrieved answer reaches or exceeds a threshold called the *confidence criterion* (C), then the

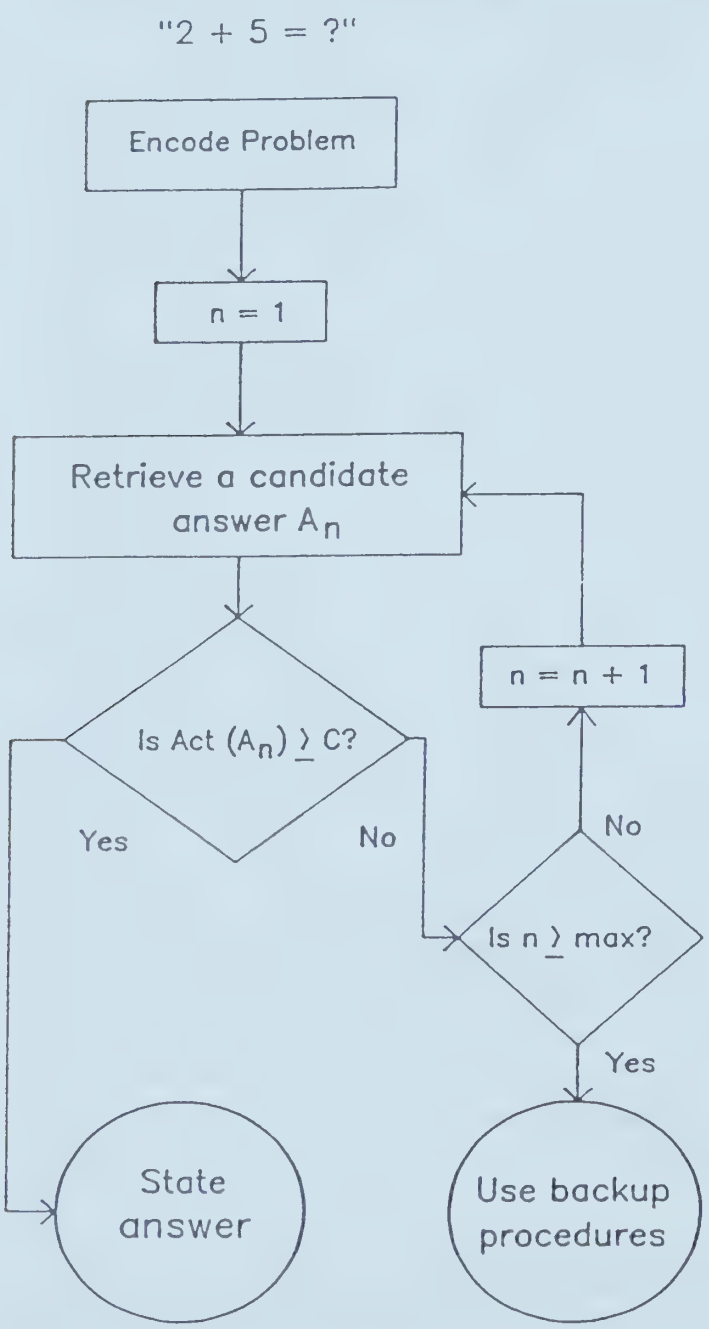


Figure 2. A simplified version of the selection process in Siegler's model of mental arithmetic. From "The information processing perspective on cognitive development in childhood and adolescence," by R. Kail & J. Bisanz, 1992, in R.J. Sternberg & C.A. Berg (Eds.), Intellectual Development *(p. 236). New York: Cambridge University Press. Copyright 1992 by Cambridge University Press. Reprinted with permission.*

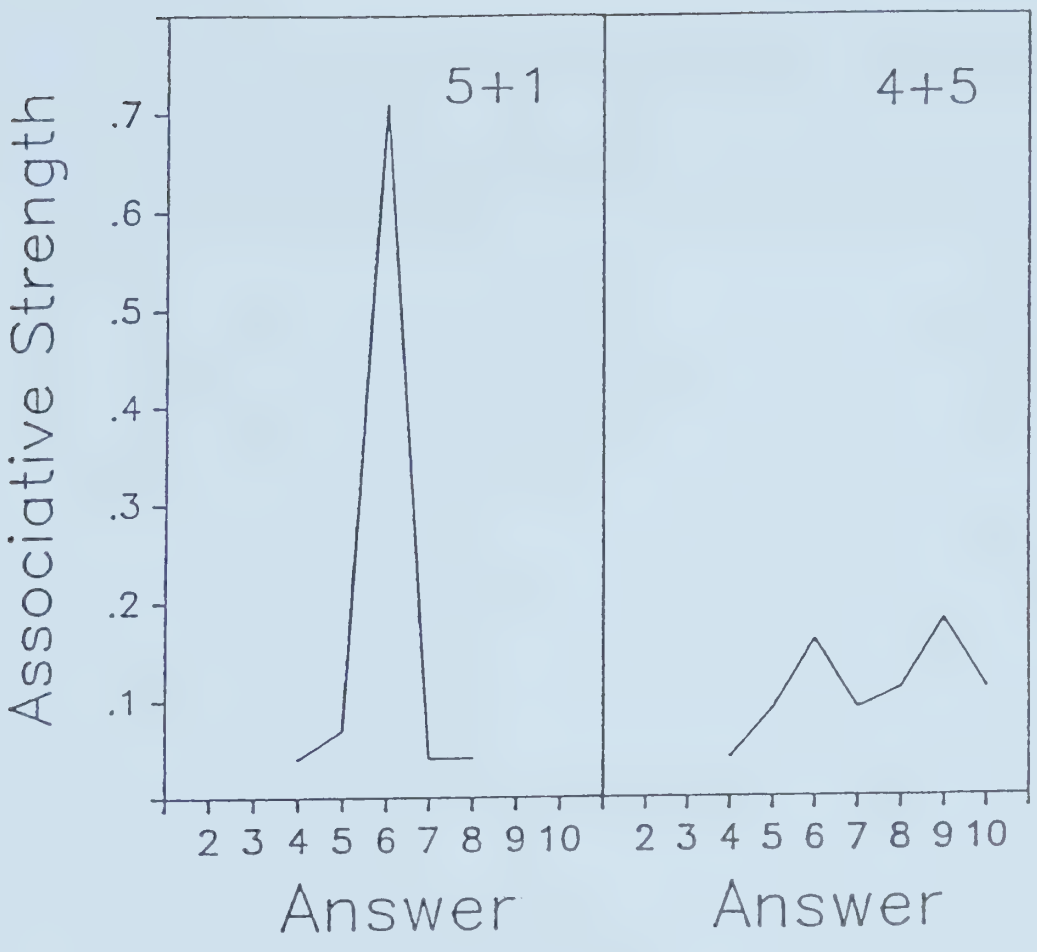


Figure 3. Distributions of associative strengths for two problems. From "Strategic and nonstrategic processing in the development of mathematical cognition" by J. Bisanz & J. LeFevre (1990) in D.F. Bjorklund (Ed.), Children's strategies: Contemporary views of cognitive development *(p. 220). Hillsdale, NJ: Lawrence Erlbaum Associates. Copyright 1990 by Lawrence Erlbaum Associates. Reprinted with permission.*

answer is stated. If the retrieved answer does not exceed the confidence criterion, then additional retrieval attempts are initiated until an answer exceeding the confidence criterion is retrieved or a limit (*max*) is reached on the number of retrievals or the time spent retrieving. If all retrieval attempts are unsuccessful, then a backup procedure is invoked to generate an answer. The success of the retrieval phase depends jointly on the distribution of associative strengths for a particular problem (i.e., how well the person knows the arithmetic fact in question) and on the level of the confidence criterion. For example, with the distributions illustrated in Figure 3 and a confidence criterion of .5, a child is likely to retrieve the answer to 5+1 and use a backup procedure to solve 4+5. With a confidence criterion of .1 a child is likely to retrieve answers to both problems, but accuracy on the second problem is likely to be low.

We hope this brief account conveys a sense of Siegler's theory of mental arithmetic and its development. Since its inception almost a decade ago (Siegler & Shrager, 1984), the theory has been successful in a number of senses (see Bisanz & LeFevre, 1990, and Siegler & Shipley, in press, for summaries). First, it helps us to account for a variety of empirical data, including correlations among solution latencies, accuracy rates, and the frequencies with which retrieval and backup procedures are used. Second, the model has been used to

generate novel and even counterintuitive predictions that have been supported empirically. Third, the model has been tested in mathematical and computer simulations that have led to interesting insights about the nature of learning mechanisms and, more generally, about cognitive development. Fourth, the model is generally compatible with more specific theories about retrieval mechanisms and problem solving in children and adults, and it has been useful for generating new research on these topics (e.g., Geary, Brown, & Samaranayake, 1991; LeFevre, Kulak, & Bisanz, 1991; LeFevre et al., 1994). Fifth, the model has considerable generality, and it has been used to identify certain similarities in development across domains (e.g., Siegler, 1986). Sixth, the theory has been revised successfully to account for additional phenomena, such as selection from among competing backup procedures (Siegler & Shipley, in press).

Finally, and of special importance for assessment, the model has proved useful for identifying insights about the nature of individual differences in performance (e.g., Geary et al., 1991; Siegler, 1988). For example, Siegler (1988) identified three groups of children who were classified as good students, not-so-good students, and perfectionists. Good students and perfectionists were more accurate than not-so-good students, but perfectionists used retrieval much less frequently than good students. These differences are readily interpretable in terms of two key parameters of Siegler's model: the peakedness of distributions and the level of the confidence criterion. Both perfectionists and good students have peaked distributions (as in the left panel of Figure 3), whereas the not-so-good students have flatter distributions. The difference between the two more successful groups is that perfectionists have higher confidence criteria than good students. The difference between good students and perfectionists is psychologically significant and potentially important for instruction, but it would have been impossible to detect with standard measures of achievement because both types of students would do well.

Understanding Arithmetic

Siegler's model concerns the computational processes used to solve arithmetic problems, but in its current form it sheds little light on the conceptual knowledge people have of arithmetic. Certainly how we solve problems is related to our understanding of arithmetic, but the nature of that relation is not very clear at present. The issue of understanding is central to many arithmetic teachers who commonly observe that their students are capable of solving simple problems when they are presented symbolically (e.g., 7-4=?) but fail miserably when required to interpret a situation arithmetically (e.g., "Erik has 7 tribbles, which is 4 more than Jason. How many does Jason have?"). Moreover, children who generate correct answers may differ considerably in their understanding of underlying arithmetic principles. Clearly the relation between what we understand about problems and how we solve those problems requires careful explication.

Developmental, cognitive, and instructional psychologists have begun to attack this issue, but progress has been difficult, in part because understanding and conceptual knowledge are often defined ambiguously or inconsistently. For example, conceptual knowledge has been described as "knowledge that is rich in relationships.... a connected web of knowledge, a network in which the

linking relationships are as prominent as the discrete pieces of information" (Hiebert & Lefevre, 1986, p. 3). As intuitively appealing as such definitions might be, they are difficult to operationalize for research and assessment purposes. Moreover, any single definition would tend to make understanding an all-or-none phenomenon and would obscure different forms of understanding that may emerge as new knowledge is acquired (Bisanz & LeFevre, 1992; Greeno, 1983).

One approach to this problem is to characterize the different forms of understanding that can be assessed and then to describe understanding in terms of performance across measures of these different forms. For example, Bisanz and LeFevre (1992) propose that contexts for assessing understanding can be organized in terms of two orthogonal aspects or dimensions, as illustrated in Table 1. The vertical aspect includes many different activities that can be used to assess forms of understanding. Consider the question of whether a child understands the principle of inversion in arithmetic (a+b-b must be equal to a). *Evaluation of procedures* refers to whether a student can recognize the validity of using a shortcut to solve problems such as 3+6-6 when that procedure is observed. These judgments help to identify whether students have some knowledge about the principle of inversion, even if they are incapable of executing an appropriate procedure or simply fail to do so spontaneously. These decisions are similar to judgments about grammaticality used by linguists and psycholinguists to determine whether individuals are sensitive to syntactic rules, whether or not those individuals are capable of stating the rule explicitly. *Application of procedures* refers to whether a student spontaneously uses a noncalculational shortcut to solve such problems as 3+6-6. Finally, *justification of procedures* is assessed by determining whether a student can describe the principle or concept that makes a procedure appropriate and legitimate, independently of whether the student uses the procedure spontaneously or recognizes the validity of the procedure correctly when observing it. Each of these "activities" is relevant to understanding, but none of them

Table 1
Contexts for Assessing Understanding

	Generality	
	Activity Narrow ——————> *Broad*	
Evaluation of procedures	Recognizing the validity of a procedure on one particular task	Making a similar judgment on a variety of related tasks
Application of procedures	Using an appropriate procedure spontaneously on one particular task	Using a similar procedure on a variety of related tasks
Justification of procedures	Describing the principle or concept that makes a procedure appropriate for one particular task	Generating a similar explanation on a variety of related tasks.

Note. From "Understanding elementary mathematics" by J. Bisanz & J. LeFevre, 1992, in J. Campbell (Ed.), *The nature and origins of mathematical skills* (p. 117). Amsterdam: Elsevier Science Publishers. Copyright 1992 by Elsevier Science Publishers. Reprinted with permission.

alone is sufficient for capturing a complete sense of what it means to understand an arithmetic principle.

The horizontal aspect of Table 1 is generality, which refers to the range of problems in which successful performance is observed. A common assumption in research on problem solving is that performance is more likely to reflect genuine understanding, as opposed to behavior acquired by rote, to the extent that it can be observed in a variety of new and/or different problems (e.g., Greeno, 1983). Carraher, Schliemann, and Carraher (1988), for example, found that construction foremen in Brazil were quite proficient at converting measurements from familiar scales on blueprints. For most of the foremen, however, performance dropped dramatically when they were required to use the same operations with unfamiliar scales, indicating that their understanding of ratio conversions was limited. In the case of understanding inversion, the question is whether students evaluate, apply, and/or justify an inversion-based shortcut on a variety of problems of different forms (e.g., b+a-b) or magnitudes (e.g., 3+6-6 versus 637+353-353).

We have two purposes in presenting Table 1. The first is simply to illustrate that cognitive, developmental, and instructional researchers have developed a variety of methods for assessing the extent to which a student "understands" arithmetic. The range of methods is impressive and informative. Second, any single measure of understanding is bound to provide an incomplete or misleading picture. An alternative is to conceive of understanding in terms of a profile in the contextual space defined by such aspects as activity and generality.

> Thus the issue for assessment of individuals is not one of deciding whether an individual has or does not have understanding, but rather one of determining the pattern of successes and failures in the contextual space. Similarly, the issue for studying acquisition is not how one group performs on a particular task as opposed to another group, but rather how understanding, or evidence for some form of understanding, "spreads" in this contextual space during the course of development or instruction. (Bisanz & LeFevre, 1992, p. 124)

Learning Environments

To this point we have discussed arithmetic as if the role of specific environmental or cultural influences is minimal. Traditionally, cognitive approaches tended to be focused on mental processes and their organization, paying little attention to the nature of the environments in which children learn. The cost of neglecting the role of the learning environment is obvious, however, and there is a growing body of research on factors related to culture (e.g., Stevenson & Stigler, 1992), schooling (Bisanz et al., in press) and home environments (Song & Ginsburg, 1987). For example, Stevenson, Lee, and Stigler (1986) initially documented striking differences in mathematics achievement among children in several countries on three continents. Many of these differences are apparent as early as grade 1 in nations where schooling at that level is universal, thus indicating that these differences are not due to selective sampling. In a second wave of research, Stevenson et al. (1990) and Stevenson, Chen, Lee, and Fuligni (1991) have begun to identify a large set of cultural and pedagogic differences among these nations that might contribute to these cross-national differences. Included are methods of instruction (e.g., the use of class discussion and

questioning), attitudes toward mathematics held by parents and teachers, and out-of-school practices (e.g., homework).

Because of constraints on space, we do not describe this research in any detail. It is important to recognize, however, that developmental and instructional research on the environments of learning mathematics is growing and making striking contributions to our understanding of acquisition. Psychologists are becoming increasingly aware that if we are to understand the development of academic skills, we need careful assessments of the environments in where those skills are acquired.

Research on the Development of Early Reading Skills

Developmental Frameworks

The need for a developmental framework for understanding how children learn to read was recognized by Chall (1979) in her seminal proposal about stages in reading acquisition. Chall's goal was to develop a Piaget-like theory of reading in which reading was seen as a form of problem solving. She outlined qualitative changes in reading as a problem solving activity over the course of schooling. Early stages were focused on relating print to speech and later stages on relating print to ideas. This shift leaves her framework open to criticisms that the proposed stages reflect changing emphases that exist in the elementary curriculum, rather than reflecting the ways in which children actually learn to read.

In contrast, information processing accounts of the development of reading have been enhanced by models of skilled reading in adults, as developed by

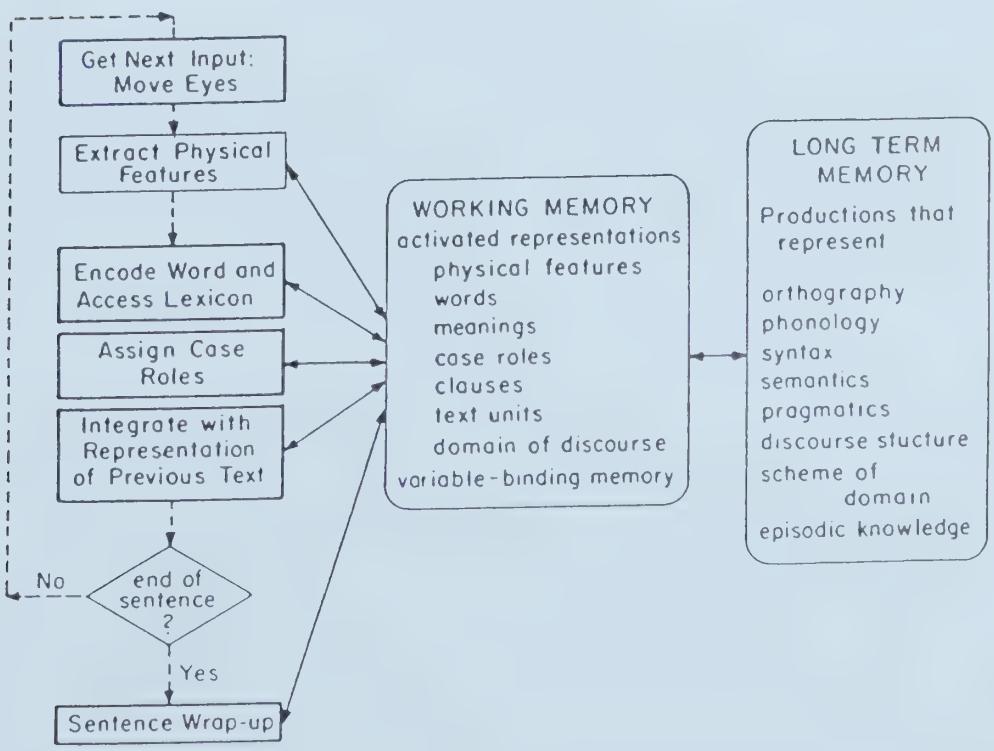


Figure 4. A schematic diagram of the major processes and structures in reading comprehension. From "A Theory of Reading: From Eye Fixations to Comprehension," by M.A. Just and P.A. Carpenter, 1980, Psychological Review, 87, p. 331. Copyright 1980 by the American Psychological Association. Reprinted with permission.

cognitive psychologists. Just and Carpenter's (1980) model of the cognitive processes that predict the eye fixations of skilled readers can serve as a vehicle to illustrate the major types of processes and structures that must be encompassed in any information processing account of skilled performance (see Figure 4). At least three questions must be addressed: What types of knowledge must the reader have in long-term memory? What is the role of working memory or limited attentional resources in performance? What are the major processes in reading and how do they interact given the reader's goal and the type of passage? From an information processing perspective, "reading" can be viewed as a family of related skills where processes could differ in temporal duration or organization or both, depending on the reader's goal and the nature of the reading materials.

For familiar words skilled readers have a large store of word codes referred to as a print lexicon by psychologists and a sight vocabulary by educators. Access to these codes involves fast and automatic retrieval, a process that uses few or no attentional resources. For novel words, skilled readers use a set of decoding or "back-up" strategies (Perfetti, 1992), including identification of subsyllabic units important in oral speech (e.g., Treiman, 1992), phonetic strategies that involve one-to-one mappings of individual phonemes to graphemes, and more sophisticated orthographic strategies that take advantage of the complexity of English spelling. These strategies all take advantage of the relations between oral language and print afforded by use of the alphabet in English. There is also some suggestion that some of the earliest and most ephemeral strategies that children use may involve the use of visual cues, treating English as if it were a logography like Chinese (e.g., Frith, 1985; Gough & Hillinger, 1980).

Children who are just learning to read in their native language begin with a relatively large oral lexicon and a nonexistent print lexicon. The critical question "What develops in early reading?" has multiple answers (see Perfetti, 1992). For familiar words it is both the number and nature of word codes, including the precision and redundancy of their representation in memory. Speed of accessing word codes also increases, but this change could well be a function of other developments. For unfamiliar words, what develops is the sophistication of back-up strategies.

Psychologists working in the area of reading and oral language hold the view that reading "piggybacks" on language in a complex way (Liberman & Shankweiler, 1979). Perfetti (1985) proposed a distinction between language processes that constitute a general reading ability and other processes of language and thought that occur in a variety of activities, including reading. Visual word recognition, for example, is a process unique to reading ability. This proposal enables one to ask whether some children could have good thinking and oral language skills but be poor readers. When asked this question, teachers generally agree some children fit this pattern.

Word recognition and thinking may be related in performance, however, because slow, inefficient retrieval of word codes is thought to take away limited resources from comprehension processes (Perfetti, 1985). Such inefficiencies can result in compensatory processing, the use of sentence-level information in the text to assist word recognition. This verbal inefficiency

should be evident in a deteriorating ability to understand the overall meaning of a text (Perfetti, 1985; Stanovich, 1980).

Many psychologists who study children's reading from an information processing perspective suggest that the word codes used in reading are also used in spelling. One difference between the two skills is that greater precision and redundancy of representation is required for spelling (Frith, 1985). Consequently, there is a body of related work on the retrieval of word codes and the development of back-up strategies in spelling (e.g., Ehri, 1987; Frith, 1985; McClure & Varnhagen, 1992; Treiman, 1992; Varnhagen & Treiman, 1993). A working hypothesis is that programs in spelling and writing can be used to facilitate reading and vice versa, through the development of more precise and redundant word codes and teaching strategies that will enhance performance on both tasks (Ehri & Wilce, 1987). Of importance to research and assessment is that the hypothesized relation between these skills opens new avenues for exploring competence.

This emerging information processing framework on the development of reading has many parallels with the work on early arithmetic and just as many implications for assessment. For developmental psychologists and assessment specialists, important tasks include determining the size and organization of a child's print lexicon, as well as examining the nature and efficiency of word codes. In addition, information about compensatory processing and the range and sophistication of back-up strategies could become important for measuring cognitive growth and achievement.

Individual Differences

Beyond this general framework for measuring the growth of all children, developmental research has begun to provide a picture of children with special strengths and weaknesses that make them more or less likely to profit from classroom instruction. For example, some children appear to have normal oral language skills just before they enter school, but they perform poorly on special tasks requiring that they segment and manipulate subsyllabic units of oral language. These children tend to score poorly on tests of reading achievement at the end of grade 1 or 2. Identification of these phonological deficits has been called a scientific success story (Stanovich, 1987).

New and provocative longitudinal research conducted by Scarborough (1990) suggests that their problem is related to a deficiency in language skills that is evident in different forms at younger ages. For instance, at age 2 these same children show deficits in syntactic complexity that are not evident later at school entry. Not surprisingly, many new studies are focused on appropriate intervention strategies. This research exemplifies the benefits of close links between basic research and assessment in areas where advances are rapid and understanding is vitally important to the welfare of children.

Similarly, research has shown that "precocious readers," those children who arrive at the school door already knowing how to read, seem to be distinguished from their peers by interest in "cracking the alphabetic code" rather than by IQ or home background (Siegler, 1991). Teachers need to take advantage of the passions of young experts to promote their intellectual growth. Assessment specialists and researchers need to identify behaviors reflecting the special interests some children may have that cause them to excel

in reading, writing, arithmetic, science, and other subject matter areas, as well as methods for motivating children who lack those special interests.

Learning Processes in Home Environments

Although developmental psychologists interested in reading often start with a focus on the mind of the child, analyses extend to understanding factors that affect learning in the various environments where children grow and learn. For example, insights into the lives of precocious readers spring from new interests among developmental researchers in documenting processes in home environments that promote literacy. Research on emergent literacy in preschoolers is focused on the naive theories of reading and writing that children develop before coming to school (e.g., Bissex, 1980) and what parents of young children do at home to promote literacy.

Developmental researchers with an interest in emergent literacy owe much to the pioneering work of educators such as Clay (1972), who observed preschoolers' exploratory behavior related to reading and writing and recognized its importance to school instruction. Other early pioneers include language researchers Ninio and Bruner (1978), who went into a home to document the development of oral language in a mother-child pair and became interested in joint book reading. Case studies and ethnographic research have been supplemented by experimental research and intervention studies (e.g., Cornell, Senechal, & Broda, 1988; Senechal & Cornell, 1993; and Whitehurst et al., 1988) focused on the types of questions parents ask their children during joint book reading and on how these questions influence memory and the acquisition of literacy skills.

Teachers' knowledge of this body of research has done a great deal to change the early elementary curriculum from a focus on "reading readiness" to an approach where children's earliest theories about reading are recognized and the need to jointly enlist the talents of the child's teachers at home and at school to facilitate learning. The challenge for educational assessment is to find ways to help teachers develop tools that will allow them to tailor their programs to meet the needs of children with particular types of naive theories or home reading experiences. Tools are also needed to help parents and teachers assess the quality of the dialogues they have with children during joint activities such as book reading. Diagnostic tools for determining the quality of a dialogue would allow parents and teachers to understand what is most effective and when they have achieved it.

Effects of Classrooms and Curricula on Cognition

Several studies have helped to make many developmental psychologists sensitive to the effects of classroom practices and curricula on psychological processes. For example, Evans and Carr (1985) conducted an exemplary study on the relation between philosophies of teaching reading (in this case, language experience versus decoding-oriented basal reader approaches to teaching reading), classroom learning activities, and the cognitive skills children acquire. They measured reading achievement and also looked for sensible patterns of positive correlations between reading and basic skills in a variety of areas, a criterion for evaluation they refer to as cognitive coherence. They found that children in language experience classrooms actually spent less time in reading

and language activities than children in traditional classrooms, a reality that was reflected in the achievement test scores. Furthermore the language experience program was not cognitively coherent in that negative correlations were found between linguistic ability and reading skill in these classrooms. This study clearly demonstrates the importance of ongoing classroom activities and the potential irrelevance of a teacher's stated philosophy to patterns of cognitive skills that children acquire.

Similarly, Perfetti, Beck, Bell, and Hughes (1987) reported data from a longitudinal study showing that the patterns of acquisition for different phonemic skills varied with the type of reading instruction children received. These studies have made it increasingly clear that, at a minimum, relevant characteristics of both subjects and classroom activities are important to consider in evaluating in reading research.

The Effect of Culture on the Study of Human Development

Finally, environments relevant to the development of reading include not only home and school, but also properties of the orthography of children's native language and the customs of their culture and subculture. For example, writing systems differ with respect to the basic linguistic units (e.g., phonemes, syllables, morphemes, words) they represent. Comparative developmental studies have begun to explore the differential effects of orthography on children's early acquisition of reading skill (e.g., Caravolas & Bruck, 1993).

Similarly, work on the literacy events in the homes of children from different communities (e.g., Heath, 1982) has gone a long way toward sensitizing developmental researchers and educators to aspects of our multicultural society that may enhance or detract from a child's opportunities to profit from classroom instruction. Moreover, comparative work on reading achievement in different cultures (e.g., Stevenson et al., 1985) has led not only to analyses of differences in home and school environments, but also to insights on differences in shared cultural values that relate to schooling.

In developmental psychology, the scope of investigation has broadened to one where each culture can be viewed as a treatment in a broad, "natural" experiment. Documenting these "treatment effects" is critical for understanding what is common about basic psychological processes and what varies with cultural context. This view of human development implies that cross-national studies of achievement must be designed and interpreted carefully so that the influence of culture on the development of academic skills can be well understood.

Implications for Assessment

Given these brief, selective summaries, we highlight several specific implications for the development of new methods of assessment.

Specific Implications

A wide range of methods now exists for developing assessments focused on identifying processes, rather than just products. It is widely acknowledged that achievement tests designed to assess products of learning, rather than processes, are inherently limited in terms of their value for teachers and for students. Decades of cognitive research have confirmed what teachers have known all along: students sometimes get the right answers for the wrong reasons, and they

sometimes get the wrong answers for the right reasons. We have tried to demonstrate that a wide range of methods exists for successfully identifying the cognitive underpinnings of performance on academic tasks. What is needed now is careful research and development aimed at adapting these methods for classroom use.

Assessments of processes and knowledge need to be closely tied to cognitive developmental theories. If the goal is to identify cognitive processes involved in children's academic performance, then an adequate theory about those processes is essential for developing assessment methods and for interpreting the results of assessment. Siegler (1988), for example, has shown how an explicit theory can provide a critical basis for diagnosing and characterizing individual differences in mental arithmetic. Theories can be be important for designing appropriate instruction to compensate for children's misunderstandings (e.g., Cummins, 1991) or for identifying developmental milestones that can serve as informative signals about a student's level of performance. The sensitivity and utility of an assessment will be proportional to the power and accuracy of the theory on which it is based; as Lewin notes, "There is nothing so practical as a good theory."

Assessments must be designed to diagnose diverse as well as modal abilities and skills. There is a tendency to describe stages or levels of competence in terms of a single element—a strategy, procedure, rule, code, or concept—that dominates cognition during a given period of time. Recent research suggests a different view: At any point in development, individuals may have a repertoire of diverse elements, and elements from this repertoire may be selected adaptively depending on external conditions or internal factors. According to some views, this diversity is not only common, it is an essential characteristic of adaptive systems (e.g., Siegler, 1989). From an instructional point of view, it may be useful to know not only what the dominant element is, but also to know (a) whether other elements are available to the child for use under certain conditions and (b) the characteristics of processing that determine selection of elements (e.g., the peakedness of distributions or the level of the confidence criterion in arithmetic). Assessment methods need to be developed to detect not only modal elements, but also to elicit the range of different elements available.

Hands-on research with children is critical for developing and selecting test items. Traditionally, test designers have attempted to designate certain items as requiring certain types of processes, based on their intuitions about solution processes. For example, according to Bloom's (Bloom, Englehar, Furst, Hill, & Krathwahl, 1956) popular taxonomy, items can be identified as requiring distinct processes such as "knowledge," "comprehension," or "application." Research with children suggests that such an enterprise is likely to be misleading: Children often solve problems with unanticipated methods, and even simple solutions typically involve an interaction among several different types of processes (Bisanz, Bisanz, & LeFevre, 1984; Gierl, 1993). To avoid misleading and adultocentric views of how problems are solved, the mind of the child must become not only a central focus for developmental psychology, but also for assessment. Items must be tested with children using a full range of meth-

ods (e.g., latencies, self-reports, and accuracy rates) to identify the range of solutions children use.

Assessments need to be designed to measure special interests, motivations, and attitudes. To fully understand the source of individual differences, assessments need to be designed to assess the extent to which children enjoy certain tasks or domains. This information could prove extremely useful to teachers who wish to adapt instruction to the particular goals and needs of individual children.

Assessments are needed of the social and cultural processes that characterize children's learning environments. Research has demonstrated important influences of people and factors in the environments where children learn, yet assessment practices often proceed as if learning environments were irrelevant or unmeasurable. Assessments of the social and cultural processes that characterize children's learning environments would not only help to provide more complete explanations of a child's performance, but they also should help teachers determine the kinds of remedial instruction that are likely to be most effective with particular children.

Epilogue

Our review of research on cognitive development and its implications for assessment would be incomplete without two additional reflections. First, we have described research on cognitive development and research on assessment as if the flow of information between these two areas were unidirectional. Such a state of affairs would be far from ideal. The flow of information should be reciprocal: the results of developmental research should inform the development of assessment methods and practices, and the development and use of assessment devices should inform and challenge research on development. Second, if this reciprocity is to be achieved, some researchers and curriculum specialists need to acquire expertise in both domains. We believe that productive interactions between assessment and cognitive development may well depend on the degree to which people in these specialities collaborate to produce individuals with the skills and insights that can arise from an interest and expertise in both domains.

Acknowledgments

The authors gratefully acknowledge support from grants provided by the Social Sciences and Humanities Research Council and by the Natural Sciences and Engineering Research Council.
Correspondence should be addressed to either author at the Centre for Research in Child Development, Department of Psychology, University of Alberta, Edmonton, AB, Canada T6G 2E9.

References

Ashcraft, M.H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review, 2*, 213-236.

Ashcraft, M.H. (1987). Children's knowledge of simple arithmetic: A developmental model and simulation. In J. Bisanz, C.J. Brainerd, & R. Kail (Eds.), *Formal methods in developmental psychology: Progress in cognitive development research* (pp. 302-338). New York: Springer-Verlag.

Ashcraft, M.H., & Fierman, B.A. (1982). Mental addition in third, fourth, and sixth graders. *Journal of Experimental Child Psychology, 33*, 216-234.

Bisanz, J., Bisanz, G.L., & LeFevre, J. (1984). Interpretation of instructions: A source of individual differences in analogical reasoning. *Intelligence, 8*, 161-177.

Bisanz, J., & LeFevre, J. (1990). Strategic and nonstrategic processing in the development of mathematical cognition. In D.F. Bjorklund (Ed.), *Children's strategies: Contemporary views of cognitive development* (pp. 213-244). Hillsdale, NJ: Erlbaum.

Bisanz, J., & LeFevre, J. (1992). Understanding elementary mathematics. In J. Campbell (Ed.), *The nature and origins of mathematical skills* (pp. 113-136). Amsterdam: Elsevier.

Bisanz, J., Morrison, F.J., & Dunn, M. (in press). Effects of age and schooling on the acquisition of elementary quantitative skill. *Developmental Psychology*.

Bissex, G.L. (1980). *GNYS at WRK*. Cambridge, MA: Harvard University Press.

Bloom, B.S., Englehar, M.D., Furst, E.J., Hill, W.H., & Krathwahl, D.R. (Eds.). (1956). *Taxonomy of educational objectives: Handbook 1. Cognitive domain*. New York: McKay.

Brown, J.S., & Burton, R.B. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*, 155-192.

Brown, J.S., & VanLehn, K. (1982). Towards a generative theory of "bugs." In T.P. Carpenter, J.M. Moser, & T.A. Romberg (Eds.), *Addition and subtraction: A cognitive perspective* (pp. 117-135). Hillsdale, NJ: Erlbaum.

Cairns, R.B. (1983). The emergence of developmental psychology. In W. Kessen (Ed.) & P.H. Mussen (Series Ed.), *Handbook of child psychology: Vol. 1. History, theory, and methods* (pp. 41-102). New York: Wiley.

Caravolas, M., & Bruck, M. (1993). The effect of oral and written language input on children's phonological awareness: A cross-linguistic study. *Journal of Experimental Child Psychology, 55*, 1-30.

Carraher, T.N., Schliemann, A.D., & Carraher, D.W. (1988). Mathematical concepts in everyday life. In G.B. Saxe & M. Gearheart (Eds.), *Children's mathematics: New directions for child development* (vol. 41, pp. 71-88). San Francisco, CA: Jossey-Bass.

Case, R. (1993, March). Central conceptual structures and their manifestations in specific task performance. In D. Kuhn (Chair), *Identifying the general in the increasingly context- and domain-specific study of cognitive development*. Symposium conducted at the meeting of the Society for Research in Child Development, New Orleans.

Chall, J. (1979). The great debate: Ten year later, with a modest proposal for reading stages. In L.B. Resnick & P.A. Weaver (Eds.), *Theory and practice of early reading Vol. 1* (pp. 29-55). Hillsdale, NJ: Erlbaum.

Clay, M.M. (1972). *Reading: The patterning of complex behavior*. Auckland, NZ: Heinemann Educational.

Cooney, J.B., & Ladd, S.F. (1992). The influence of verbal protocol methods on children's mental computation. *Learning and individual differences, 4*, 237-257.

Cornell, E.H., Senechal, M., & Broda, L.S. (1988). Recall of picture books by 3-year-old children: Repetition and testing effects in joint reading activities. *Journal of Educational Psychology, 80*, 537-542.

Cummins, D.D. (1991). Children's interpretation of arithmetic word problems. *Cognition and Instruction, 8*, 261-289.

Ehri, L.C. (1987). Learning to read and spell words. *Journal of Reading Behavior, 19*, 5-31.

Ehri, L.C., & Wilce, L.S. (1987). Does learning to spell help beginning readers learn to read words? *Reading Research Quarterly, 18*, 47-65.

Evans, M.E., & Carr, T.H. (1985). Cognitive abilities, conditions of learning, and the early development of reading skill. *Reading Research Quarterly, 20*, 327-350.

Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. E. Patterson, J.C. Marshall, & M. Coltheart (Eds.), *Surface dyslexia* (pp. 301-330). Hillsdale, NJ: Erlbaum.

Geary, D.C., Brown, S., & Samaranayake, V.A. (1991). Cognitive addition: A short longitudinal study of strategy choice and speed-of-processing differences in normal and mathematically disabled children. *Developmental Psychology, 27*, 787-797.

Geary, D.C., & Wiley, J.G. (1991). Cognitive addition: Strategy choice and speed-of-processing differences in young and elderly adults. *Psychology and Aging, 6*, 474-483.

Gierl, M.J. (1993). *Evaluating Bloom's cognitive levels in the table of specifications for a grade 6 mathematics achievement test*. Unpublished master's thesis, University of Alberta.

Ginsburg, H.P. (1986). Academic diagnosis: Contributions from developmental psychology. In J. Valsiner (Ed.), *The individual subject and scientific psychology* (pp. 235-260). New York: Plenum.

Gough, P.B. & Hillinger, M.L. (1980). Learning to read: An unnatural act. *Bulletin of the Orton Society, 30*, 180-196.

Greeno, J.G. (1983). Forms of understanding in mathematical problem solving. In S.G. Paris, G.M. Olson, & H.W. Stevenson (Eds.), *Learning and motivation in the classroom* (pp. 83-111). Hillsdale, NJ: Lawrence Erlbaum Associates.

Heath, S.B. (1982). What no bedtime story means: Narrative skills at home and school. *Language and Society, 11*, 49-75.

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 1-27). Hillsdale, NJ: Erlbaum.

Just, M.A., & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*, 329-354.

Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. *Psychological Bulletin, 109*, 490-501.

Kail, R., & Bisanz, J. (1982). Cognitive strategies. In C.R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 229-255). New York: Academic Press.

Kail, R., & Bisanz, J. (1992). The information-processing perspective on cognitive development in childhood and adolescence. In R.J. Sternberg & C.A. Berg (Eds.), *Intellectual development* (pp. 229-260). New York: Cambridge University Press.

Kail, R.V., Jr., & Hagen, J.W. (Eds.) (1977). *Perspectives on the development of memory and cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

LeFevre, J., Kulak, A.G., & Bisanz, J. (1991). Individual differences and developmental change in the associative relations among numbers. *Journal of Experimental Child Psychology, 52*, 256-274.

LeFevre, J., Sadesky, G., & Bisanz, J. (1994). *The role of strategy selection in mental addition: Reassessing the problem-size effect in adults*. Unpublished manuscript.

Liberman, I.Y., & Shankweiler, D. (1979). Speech, the alphabet, and teaching to read. In L.B. Resnick & P.A. Weaver (Eds.), *Theory and practice of early reading* (vol. 2, pp. 109-132). Hillsdale, NJ: Erlbaum.

McClure, K., & Varnhagen, C.K. (1992). *Do children spell the way they think they do? Predicted versus reported strategy use by early elementary school children*. Paper presented at the meetings of the American Educational Research Association, San Francisco.

Ninio, A., & Bruner, J. (1978). The achievement and antecedents of labeling. *Journal of Child Language, 5*, 1-14.

Pachella, R.G. (1974). The interpretation of reaction time in information- processing research. In B.H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp. 41-82). Hillsdale, NJ: Erlbaum.

Perfetti, C.A. (1985). *Reading ability*. New York: Oxford University Press.

Perfetti, C.A. (1992). The representation problem in reading acquisition. In P.B. Gough, Linnea C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 145-174). Hillsdale, NJ: Erlbaum.

Perfetti, C.A., Beck, I., Bell, L.C., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. *Merrill-Palmer Quarterly, 33*, 283-319.

Royer, J.M., Cisero, C.A., & Carlo, M.S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63*, 201-243.

Russo, J.E., Johnson, E.J., & Stephens, D.L. (1989). The validity of verbal protocols. *Memory and Cognition, 17*, 759-769.

Scarborough, H.S. (1990). Very early language deficits in dyslexic children. *Child Development, 61*, 1728-1743.

Senechal, M., & Cornell, E.H. (1993). Vocabulary acquisition through shared reading experiences. *Reading Research Quarterly, 28*, 361-374.

Siegler, R.S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8*, 481-520.

Siegler, R.S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46* (Whole No. 189).

Siegler, R.S. (1986). Unities across domains in children's strategy choices. In M. Perlmutter (Ed.), *Minnesota symposium on child development* (pp. 1-48). Hillsdale, NJ: Erlbaum.

Siegler, R.S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116*, 250-264.

Siegler, R.S. (1988). Individual differences in strategy choices: Good students, not-so-good students, and perfectionists. *Child Development, 59*, 833-851.

Siegler, R.S. (1989). Mechanisms of cognitive development. *Annual Review of Psychology, 40*, 353-379.

Siegler, R. (1991). *Children's thinking* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Siegler, R.S., & Robinson, M. (1982). The development of numerical understanding. In H.W. Reese & L.P. Lipsitt (Eds.), *Advances in child development and behavior* (vol. 16, pp. 241-312). New York: Academic Press.

Siegler, R.S., & Shipley, C. (in press). Variation, selection, and cognitive change. In G. Halford & T. Simon (Eds.), *Developing cognitive competence: New approaches to process modeling*. Hillsdale, NJ: Erlbaum.

Siegler, R.S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do? In C. Sophian (Ed.), *Origins of cognitive skills* (pp. 229-293). Hillsdale, NJ: Erlbaum.

Simon, H.A. (1962). An information processing theory of intellectual development. *Monographs of the Society for Research in Child Development, 27* (Serial No. 82).

Song, M.J., & Ginsburg, H.P. (1987). The development of informal and formal mathematical thinking in Korean and U.S. children. *Child Development, 58,* 1286-1296.

Stanovich, K.E. (1980). Toward an interactive-compensatory model of individual differences in the development of fluency in reading. *Reading Research Quarterly, 16,* 32-71.

Stanovich, K.E. (1987). Introduction. *Merrill-Palmer Quarterly, 33*(3).

Stevenson, H.W. (1972). *Children's learning*. New York: Appleton-Century-Crofts.

Stevenson, H.W., Chen, C., Lee, S.Y., & Fuligni, A.J. (1991). Schooling, culture, and cognitive development. In L. Okagaki & R.J. Sternberg (Eds.), *Directors of development: Influences on the development of children's thinking* (pp. 243-268). Hillsdale, NJ: Erlbaum.

Stevenson, H.W., Lee, S.Y., Chen, C., Lummis, M., Stigler, J., Fan, L., & Ge, F. (1990). Mathematics achievement of children in China and the United States. *Child Development, 61,* 1053-1066.

Stevenson, H.W., Lee, S.Y., & Stigler, J.W. (1986). Mathematics achievement of Chinese, Japanese, and American children. *Science, 231,* 693-699.

Stevenson, H.W., & Stigler, J.W. (1992). *The learning gap*. New York: Summit Books.

Stevenson, H.W., Stigler, J.W., Lee, S.Y., Lucker, G.W., Kitamura, S., & Chen-chin, H. (1985). Cognitive performance and academic achievement of Japanese, Chinese, and American children. *Child Development, 56,* 718-734.

Thibodeau, M.H. (1993). *Automaticity of multiplication facts and procedure use for multiplication problems: Evidence for ADAMM*. Unpublished master's thesis, University of Alberta.

Treiman, R. (1992). The role of intrasyllabic units in learning to read and spell. In P.B. Gough, L.C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 65-106). Hillsdale, NJ: Erlbaum.

Varnhagen, C.K., & Treiman, R. (1993). *On-line investigation of children's spelling*. Paper presented at the meetings of the American Educational Research Association, Atlanta.

Werner, H. (1937). Process and achievement. *Harvard Educational Review, 7,* 353-368.

Wertsch, J.V., & Kanner, B.G. (1992). A sociocultural approach to intellectual development. In R.J. Sternberg & C. Berg (Eds.), *Intellectual development* (pp. 328-349). New York: Cambridge University Press.

Whitehurst, G.J., Falco, F.L., Lonigan, C.J., Fischel, J.E., DeBaryshe, B.D., Valdez-Menchaca, M.C., & Caulfield, M. (1988). Accelerating language development through picture book reading. *Developmental Psychology, 24,* 522-559.

The Alberta Journal of Educational Research *Vol. XL, No. 2, June 1994, 147-161*

John R. Kirby

and

Rosamund A. Woodhouse
Queen's University

Measuring and Predicting Depth of Processing in Learning

Recent research regarding depth of processing in student learning is reviewed. The nature of processing depth is described as a dispositional construct and as a characteristic of task-specific strategies. Efforts to measure processing depth, as shown in students' text summaries, are reviewed. Measures include content scores (based on a weighted sum of propositions included), depth ratings, and analyses of subjects' verbalizations during study. The individual difference and task variables associated with processing depth are considered, and a number of experimental studies reviewed. The article concludes by considering several ways in which depth may be measured in educational assessment.

Dans cet article, on examine les recherches récentes vis-à-vis le degré de l'approfondissement de la matière dans le contexte de l'apprentissage chez l'élève. La nature de l'approfondissement de la matière est considérée comme étant une construction intellectuelle et dispositionnelle et comme une caractéristique des stratégies centrées sur des tâches et des habiletés spécifiques chez l'élève. On examine aussi dans le contexte d'analyses sommaires des élèves comment mesurer le degré de l'approfondissement de la matière. Les éléments qui aideront à déterminer ce degré seront le contenu des scores (basé sur la somme totale de ce qui est proposé), l'évaluation du degré de l'approfondissement de la matière, et les analyses des expressions verbales qu'utilisent les élèves pendant l'étude. On considère aussi la différence individuelle ainsi que les variables de chaque tâche associées au degré de l'approfondissement de la matière. Certaines études expérimentales qui traitent de ce sujet sont examinées. On conclu l'article en considérant différentes façons permettant de déterminer le degré de l'approfondissement de la matière quand il s'agit d'évaluation en éducation.

It has long been argued, in both education and psychology, that *depth of processing* is valuable in learning. Although definitions have varied or been lacking, there is general agreement that characteristics of greater depth include: focus on main ideas rather than details, greater connection of new information with prior knowledge, less reliance on verbatim recall, reordering or drawing connections between disparate sections of a text, drawing inferences or general principles, and ability to apply learning to a novel task. These characteristics have *going beyond the information given* as their basic feature. Depth of process-

---

John Kirby is a professor of education. After his doctoral work in educational psychology at the University of Alberta, he worked for 11 years in the Faculty of Education at the University of Newcastle, NSW, Australia. His research interests include learning and studying strategies, the cognitive processes of comprehension, and learning disabilities.
Rosamund Woodhouse is a research associate at Queen's University, where she completed her doctorate in psychology in 1990. Her research interests include learning and studying strategies and the cognitive processes of bilingualism.

ing is thought to contribute to the quality of learning by enhancing recall of information, increasing the level of importance of the information recalled, and allowing the learner to generalize learning.

The basic notion of depth of processing is that much "new" information is relatively raw, uninterpreted, and unintegrated, and needs to be transformed, interpreted, and related to other information before it can be used optimally. This begs the question of how the information is to be used, that is, what the task is. In Western education, much lip service and some serious attention is paid to the proposition that learning and teaching should focus on generalizable principles, problem solving, and understanding. Teachers and students alike show disrespect for learning and teaching that emphasize "just the facts," are not applied to "real" problems, are "low level," or require "regurgitation." In spite of these espoused beliefs, much teaching and learning is shallow, and there is legitimate concern that this is the result of evaluation practices and perceptions of them.

If beliefs about the value of depth are to be tested, and, if found to be valid, applied, we need to examine what depth is, how it can be measured, and what factors encourage it. The purposes of this article are to (a) examine the concept of depth of processing, (b) develop criteria for measuring depth of processing, (c) consider the personal and situational factors that predict depth of processing, and (d) review recent studies of learning from text that bear on these issues. This leads to discussion of methods of addressing depth of processing in educational assessment, and to the implications of a shift to this goal of assessment.

Depth of Processing

Many of our notions of depth originate in memory research in experimental psychology. In 1972, Craik and Lockhart proposed a levels of processing frame-

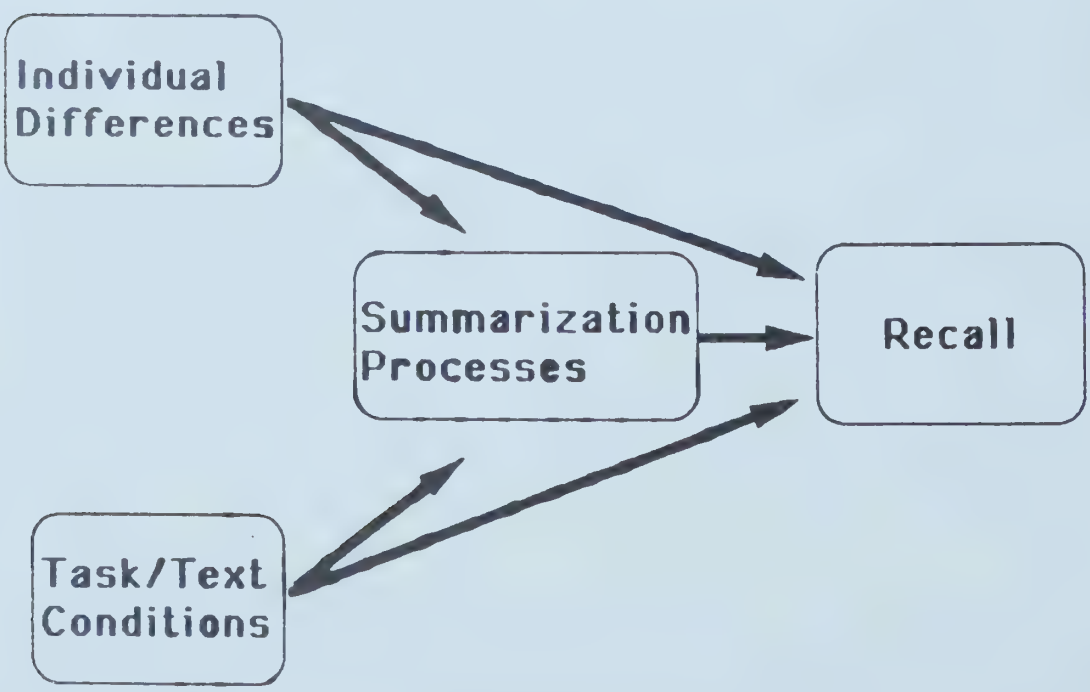


Figure 1. Model showing effects of individual differences, task factors, and study processes on recall.

work for understanding memory, where the different levels represented degrees of depth. The three levels so distinguished, the physical, phonemic, and semantic, were greatly different from each other. Only the semantic level is critical to the focus of this article, meaningful student learning. We suggest that there are degrees of depth in semantic processing that are associated with both the quantity and quality of what is remembered.

Consider Figure 1, adapted from similar figures by Biggs (1994), Duncan and Biddle (1974) and Biggs and Kirby (1984). This path diagram shows that individual differences and situational (task) factors affect processing during learning, and that all three of these affect subsequent learning outcomes. Depth of processing may be associated with each of these four components. For example, students may vary in the degree to which they are inclined toward deeper processing (an individual difference variable), which may in turn be due to other individual differences (in prior knowledge, verbal ability, etc.). As the memory researchers have shown, situations exert considerable influence over processing; task manipulations such as insertion of deeper questions, or provision of an advance organizer, are intended to increase depth of processing. Depth of processing may also be seen as something controlled by the individual student's motives and strategies at a particular time, whether during study (the middle box of Figure 1) or at performance (the rightmost box). Each of these components of depth poses problems for measurement. We review each of the components in turn, describing some recent research results, and indicating what the measurement issues are.

Disposition Toward Depth of Processing

Depth of processing has been considered as a disposition, as an individual difference regarding the likelihood of engaging in deeper processing (e.g., Biggs, 1987; Entwistle, 1988). In this sense it has been measured either by questionnaire or interview, and has been seen as a stable, transsituational characteristic, not unlike a personality trait or cognitive style. Biggs (1987) defined three dimensions of students' "approaches to learning": Deep, Surface, and Achieving, each having a motive and a strategy component. The Deep Approach to learning is characterized by intrinsic motivation and meaningful learning strategies, the Surface Approach by extrinsic motivation and rote learning strategies, and the Achieving Approach by achievement motivation and task-oriented, organized strategies. Students may have any combination of these characteristics; thus, for example, a student may show high levels of both Deep and Surface Approaches. This is not necessarily inconsistent, as some learning may require both; in fact, a combination of Deep and Surface is typical of university Science students.

There is considerable evidence that these approaches to learning variables are related to student learning, but the effects are far from powerful. In general, the surface approach is negatively related to success, the deep and achieving approaches positively related. Overall, the approach variables usually account for less than 15% of the performance variance. Biggs and Rihn (1984) found that university students asking for help at a learning assistance center had a predominantly surface approach, which shifted toward a deeper approach after completion of the center's program.

Although no one would seriously suggest that these approach measures be used to assess learning outcomes, they may be worth considering in conjunction with other factors. For example, Lai and Biggs (1994) have shown that mastery learning conditions favor, and are preferred by, high school students who report a surface approach to learning; although mastery learning improved group achievement, this was due to improvements by the surface learners at the expense of a deterioration in the performance of the deeper learners. As we see in the next section, students' approaches to learning also may determine how they react to instruction designed to increase depth of processing. These measures may have some benefit in helping students understand their own nature as learners. The validity of the transsituational nature of these approach variables should also be further examined (Biggs, 1993).

It is important to determine the antecedents of these approach variables. For example, a deep approach may require greater verbal ability and available working memory. A deep approach may only develop in an encouraging context; alternatively, the surface approach may develop as a response to particular experiences. The existing research tends to deal with approaches as already existing, rather than inquiring how they developed (see Biggs, 1987, 1993). If different educational tasks require approaches that differ in the relative amounts of deep and surface approaches, then optimal performance should be shown by students who understand these contingencies and can adapt their approach to the task at hand.

Conditions Affecting Depth of Processing

Depth of processing is a construct that is far from easy to measure unambiguously. A good experimental strategy for dealing with this is to measure the results of the process rather than attempting to measure the process itself. If the hypothesis is that greater depth will result in enhanced learning, design two treatments that differ in the depth they encourage, and observe the outcomes. There have been numerous attempts to do this with respect to constructs related to depth of processing. For the purposes of the present article, we describe two techniques, text absent summarization and poorly structured text.

Text absent summarization. Many instructors fear that students are too passive when they study. This passivity allows a form of comprehension that does not require much depth of processing and may result in shallow or fragile learning. Hidi and Anderson (1986) proposed that students would be forced to become more active if they were asked to summarize the text they were studying, but only after the text had been removed. Text absent summarization should encourage students, or at least some of them, to be more integrative and constructive during study, which should result in greater subsequent recall of important information from the text. We have examined this technique in a number of studies (Kirby & Pedwell, 1991; Stein & Kirby, 1992; Woodhouse & Kirby, 1991).

In general, we have not found across-the-board improvements in learning as a result of text absent summarization, in comparison with more traditional text present summarization. Instead, text absent summarization appears to benefit certain students, while adversely affecting the performance of others. The students who benefit are those of greater initial ability, those who write better summaries, or those who report a deep approach to learning. Text absent

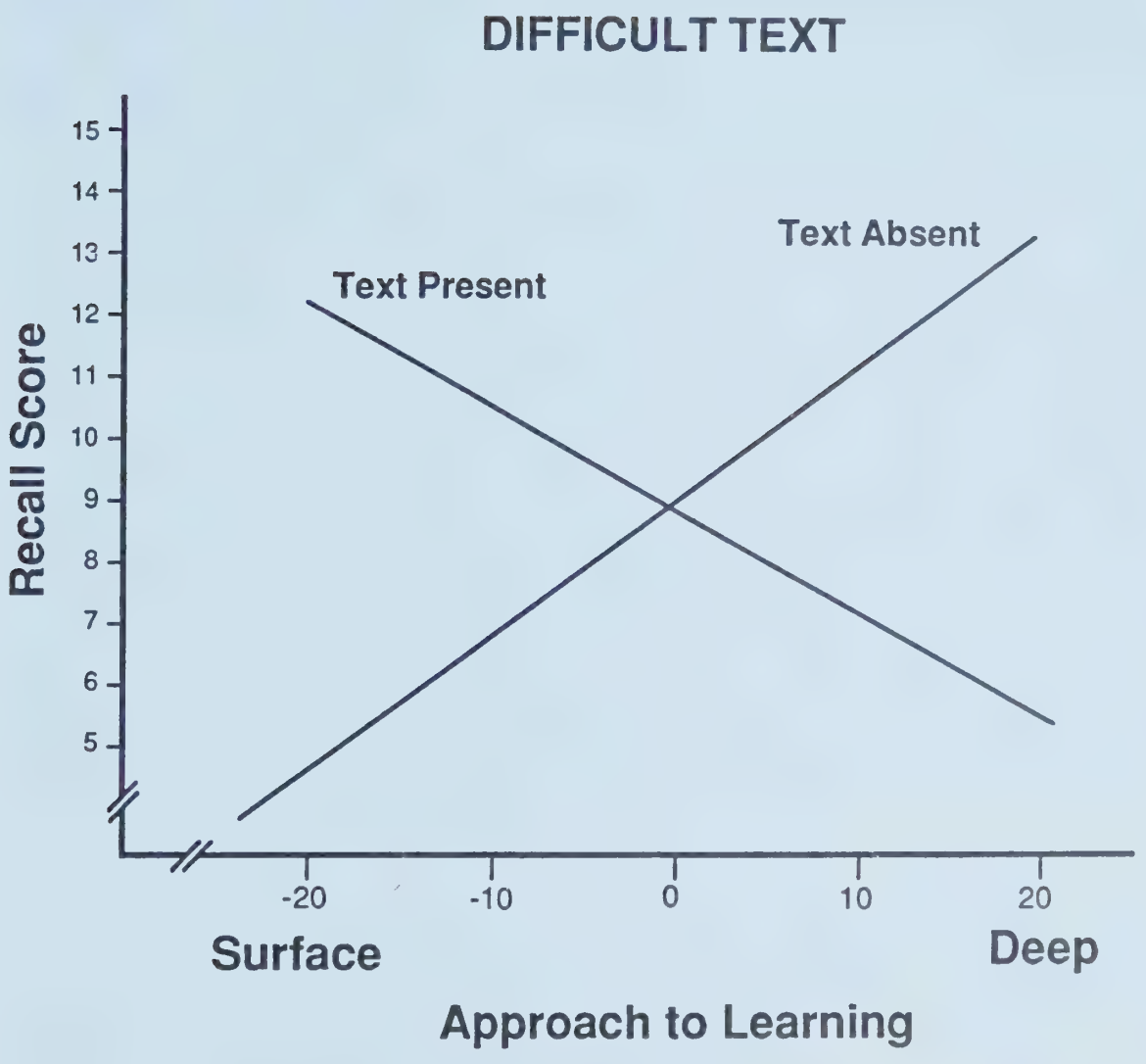


Figure 2. Regression lines illustrating interaction between task condition and students' approaches to learning (adapted from Kirby & Pedwell, 1991).

summarization is essentially a more onerous or stressful form of study: if you can cope with its demands, it helps; otherwise it may be counterproductive.

Kirby and Pedwell (1991) found an interaction between summarization condition (text absent/present) and university students' reported approaches to learning (defined as the relative tendency toward a deep or surface approach, calculated by subtracting the surface score from the deep score, see Figure 2). In the text absent condition, recall was positively associated with the relative depth of the students' habitual approach to learning; in the text present condition, this relationship was negative.

Stein and Kirby (1992) employed the text absent/present technique with grade 7 students. They found that text absent subjects wrote deeper summaries and that those who did remembered more of the important information in the text. We return to the analysis of summary quality after considering a second technique for enhancing depth of processing.

Poorly structured text. A second comparison has been made between text that is well structured (user-friendly or considerate) as opposed to text that is poorly structured (less friendly or considerate). Although very poorly structured text (sentences or words in random order) would be impossible to com-

prehend, much less process deeply, poorly structured text that is merely difficult (i.e., still comprehensible but at greater effort) has been shown to enhance processing, at least in some subjects (Kintsch, 1990; Simpson, 1992). Our interpretation of this result is that poorly structured text leads some subjects to more active and deeper processing.

In a recent study (Simpson, 1992; Woodhouse, Kirby, Simpson, & Hadwin, 1992), we had university student subjects study one of two versions of a text, one in which the sequence of main ideas was maintained and one in which it was disrupted. Subjects were asked to summarize the text and two days later to do a free and a probed recall of it. Although the poorly structured group wrote less extensive summaries than the well-structured group, they did not differ in subsequent free recall. By itself, this is only weak support for the effects of poorly structured text. However, two other results offer stronger support. First, an interaction appeared between experimental condition and subjects' efficiency of access to working memory (as measured by a speed of word naming test); this interaction showed that poorly structured text yielded superior free recalls for subjects with more efficient working memory (see Figure 3). Presumably, more efficient working memory processes offered these subjects the opportunity to perform deeper processing while reading and summarizing the text. This was supported by the second finding, which was that the quality of summary

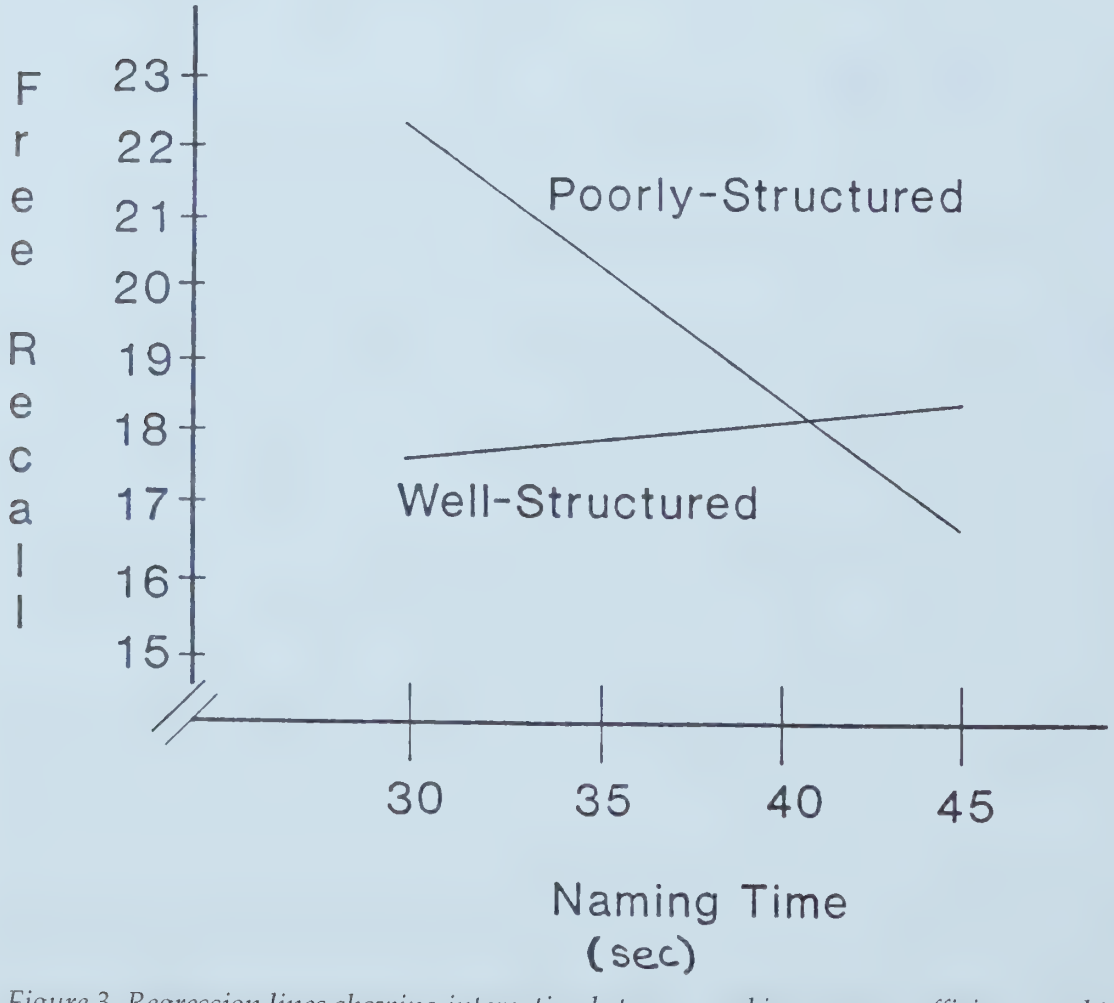


Figure 3. Regression lines showing interaction between working memory efficiency and text structure (adapted from Simpson, 1992).

written was a powerful predictor of recall for the poorly structured group. This latter effect eliminated the working memory effect, suggesting that greater working memory efficiency allowed deeper processing during study, which resulted first in a better summary, and second in better recall. (We examine the summary quality measures in the following section.) These effects appeared for free recall, but not for probed recall. In free recall, subjects must rely on their knowledge structures to "construct" a response, whereas in probed recall the question acts as a cue to trigger a response; in the latter case depth of processing is not as critical.

In a subsequent study (Peters, 1993), we replicated the working memory and text structure effects, and showed that instruction in deeper processing during study/summarization improved subjects' performance. We interpret this to mean that task conditions can have a powerful effect on the quality of processing employed by students during study, and that there is considerable scope for improvement. We return to these issues later in the article.

More generally, these task-imposed differences in processing offer several potential areas of application. First, they force us to recognize that tasks do differ in the processing they elicit and that these differences may not be constant across subjects. To accurately assess depth of processing, a range of task formats will need to be used, and the notion of a fixed "ability" for depth of processing may either be mistaken or in need of stated contextual conditions. Second, they suggest the possibility of a form of dynamic testing, in which some baseline level of processing is compared with that obtained in a task format that encourages deeper processing. Third, these different tasks may be valuable as training tasks to improve depth of processing. There is little danger of teaching to the test here, as the greater depth in study is exactly what we are attempting to encourage.

Depth of Processing During Study

To this point we have considered the first two boxes in Figure 1; essentially this is the same as aptitude-treatment interaction research, relating outcomes to some interaction between stable individual difference variables and environmental factors (more generally, Person-Task interactions). The model becomes more complex, and may gain more explanatory potential, with the addition of the Study Process component (Person-Task-Process interactions). Few doubt the importance of cognitive processes during learning; the question, rather, is how to measure these processes. We have used a variety of methods for measuring these processes during summarization.

Summarization seems an ideal context in which to examine study processes, because the student leaves a visible record of what he or she thinks is important. There are of course limitations. Summarization may be unfamiliar to some students, and some may not choose to write down what they think is important. Summarization may change the nature of learning, so that what we observe is not what would otherwise have occurred. Figure 4 presents a schematic outline of the cognitive processes involved in summarization. These processes are in three groups: on the left, those related to the student's comprehension of the text; in the middle, the summarization processes that depend on comprehension; and on the right, a description of the study product (the summary) that emerges. Comprehension processes are clearly relevant to

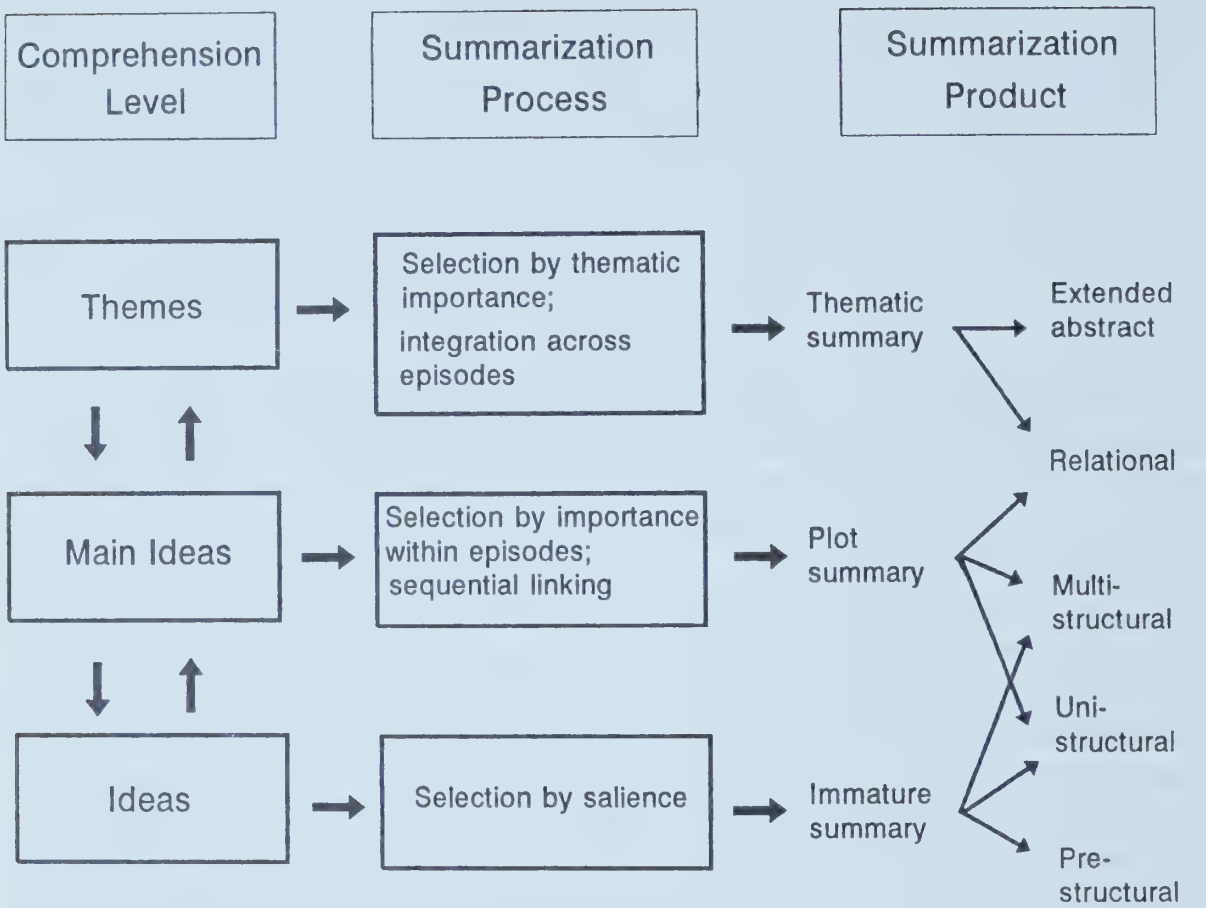


Figure 4. Cognitive processes in summarization.

learning from text and may be characterized in terms of the level at which they operate (Kirby, 1988). Three levels are shown, those relating to ideas (propositions), main ideas (macropropositions), and themes. Comprehension at the idea level is able to deal with individual ideas, as perhaps represented in a sentence, but not able to integrate across ideas or sentences or to recognize the relative importance of ideas. Comprehension at the main idea level can integrate across ideas or sentences and recognize relative importance of ideas, but does not go beyond the text to recognize the relative importance of main ideas. Comprehension at the thematic level does go beyond the text, applying an abstract frame of reference to its interpretation (see Kirby, 1988 and 1991 for more on this).

Summarization processes consist essentially of selection, linking, and construction. Selection is the means by which ideas are identified for inclusion in the summary. Linking is the process by which these ideas are connected, and construction is brought to bear when no appropriate idea can be selected or when an interpretation is made. These processes depend critically on the level of comprehension operating. If comprehension is operating at the idea level, then selection can only be on the basis of salience ("this is interesting") or by a rote rule ("copy the first sentence of each paragraph"); linking and construction hardly operate at all. If comprehension is at the level of main ideas, then selection can operate quite effectively in main idea units (roughly, paragraphs), but poorly across them. Linking works to form a sequential structure, essentially a plot line. Construction can operate in main idea units to form ideas, but

does not operate effectively between main idea units. If comprehension is at the thematic level, then selection, linking, and construction can operate effectively, with the lower levels interpreted in terms of the thematic level. Whereas main idea comprehension yields summaries that attempt to pay equal attention to each section or episode (until fatigue sets in or the paper runs out), thematic comprehension allows summaries that pay disproportionate attention to certain main ideas or even construct new main ideas.

These summary processes produce summaries of different quality. Idea level comprehension and selection by salience are unlikely to produce anything more than an immature summary. Main idea comprehension is more likely to result in a plot summary, an episode-by-episode recounting of the text. Thematic comprehension should result in a higher level, more abstract summary, which may in fact omit many of the details included at lower levels. These types of summaries can be characterized in terms of Biggs and Collis' (1982, 1989) SOLO (Structure of Observed Learning Outcomes) Taxonomy: immature summaries will either be prestructural (they miss the point), unistructural (they pick out one point), or at best multistructural (they pick out several points without integrating them). Plot summaries are essentially multistructural, but in some texts they may succeed in being relational (presenting a coherent framework within the bounds of the text). Thematic summaries will be at least relational, and should be able to reach Biggs and Collis' highest level, extended abstract (adoption of a frame of reference that is independent of the text). We return to the SOLO Taxonomy later, as this is one method of assessing summary depth.

We have used three different approaches to measuring these processes during summarization: measures of summary content and summary depth, and analyses of subjects' verbalizations while studying and summarizing. We consider each of these in turn, and provide some illustrative results.

Summary content. The summary content scores attempt to provide a measure of the information contained in the summary, weighted for importance. To do this, we analyze the text into idea units and form these idea units into a hierarchical structure. The idea units are referred to as propositions, but tend to be larger units than the propositions defined by others (e.g., Fredericksen, 1975; Kintsch & van Dijk, 1978). Essentially, an idea unit corresponds to a simple sentence or to an inferred idea similar in scope to a sentence. The idea units (some directly from the text and some inferred) are organized into a structure with three or more levels: typically we employ the levels of themes, main ideas, important ideas, and unimportant ideas, although in some texts it is not possible to identify a thematic level. These levels correspond, of course, to the different levels of comprehension referred to above. Scores may consist of the number of units at each level, or of a total score, formed by weighting the levels (usually 5 for themes, 3 for main ideas, and 1 for important ideas). The nature of the text determines the number of ideas at each level (the text structure), and the difficulty in agreeing on that structure. The scores that a subject obtains should reflect that subject's cognitive structure for the text information. Content scores are used to score summaries or free recalls of the entire text or can be adapted to score probed recall (short answer) questions.

To illustrate the use of content scores, Figure 5 shows an interaction obtained by Stein and Kirby (1992) in their study of children's learning from text in text absent and present summarization. Summary content scores are significantly and positively related to recall in the absent condition, but not in the present condition. The summary content score appears to be a good index of depth of processing in the former condition, but not in the latter. In the text absent condition, subjects were able to copy directly from the text, so their high content scores do not necessarily reflect greater depth of processing.

Summary depth. Our second approach to measuring depth in a summary relies on rating the complexity of the structure or processes underlying the summary. This is clearly more subjective. In some cases we have felt that it yields a more appropriate index of a summary's depth, although this is not always the case. The depth rating may be most appropriate when students have access to the original text during summarization and copy large amounts of text into their summary—this can yield an unrealistically high content score, although often a disastrously low depth rating.

In our most recent work (e.g., Fuller, 1992), we have used the following scale: 1 point: no coherence; 2 points: limited clusters of coherence; 3 points: linear coherence (following the sequence of the text); 4 points: relational framework, within the bounds of the text; and 5 points: a relational framework that extends beyond the text. These depth ratings focus on the structure and coherence of the summary, and thus resemble Biggs and Collis' (1982) SOLO scores.

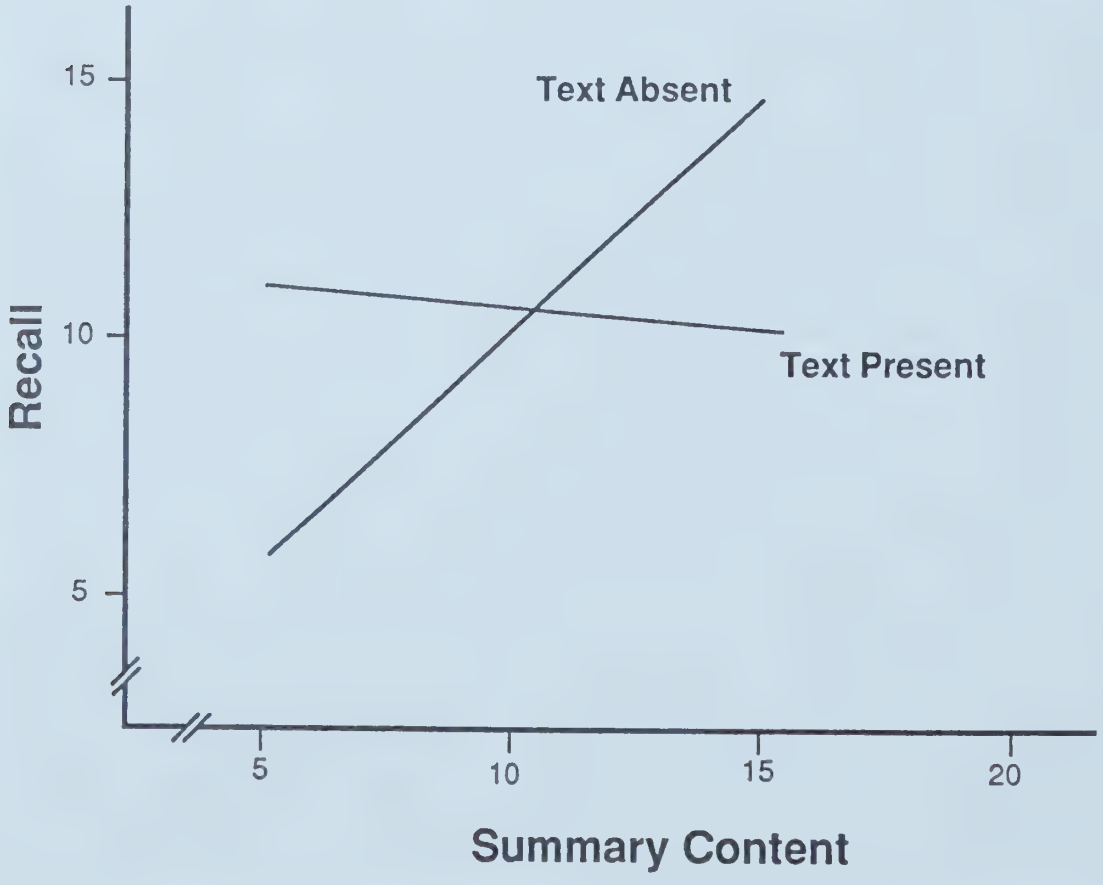


Figure 5. Regression lines showing interaction between task conditions and summary content (adapted from Stein & Kirby, 1992).

Figure 6 shows regression lines obtained in a recent study (Fuller, 1992), relating the depth of processing shown in a summary to recall in two conditions. The conditions differed in the nature of the questions subjects were asked to answer while they studied the text, either relatively surface questions (fact-oriented) or relatively deep questions (oriented toward main ideas, inferences, and interpretations). The deep questions were intended to increase depth of processing. As can be seen in Figure 6, rated depth of processing was not related to recall in the deep questions group; presumably these subjects had already carried out deeper processing during their study, and did not have to do so in their summary. For the surface questions group, however, rated depth of processing accounted for a significant proportion of the variance in recall. Having not been led to depth during study, these subjects needed to attain it in their summaries.

In other work (e.g., Stein & Kirby, 1992), especially with younger readers, we have relied on estimates of the amount of transformation required to go from the text to the summary, as indexed by the number of connectives used, reorderings, lack of copying, and whether or not the writer provided an independent perspective on the text ("What the author was trying to say was ..."). Stein and Kirby found that rated summary depth was significantly related to recall in both text absent and present conditions.

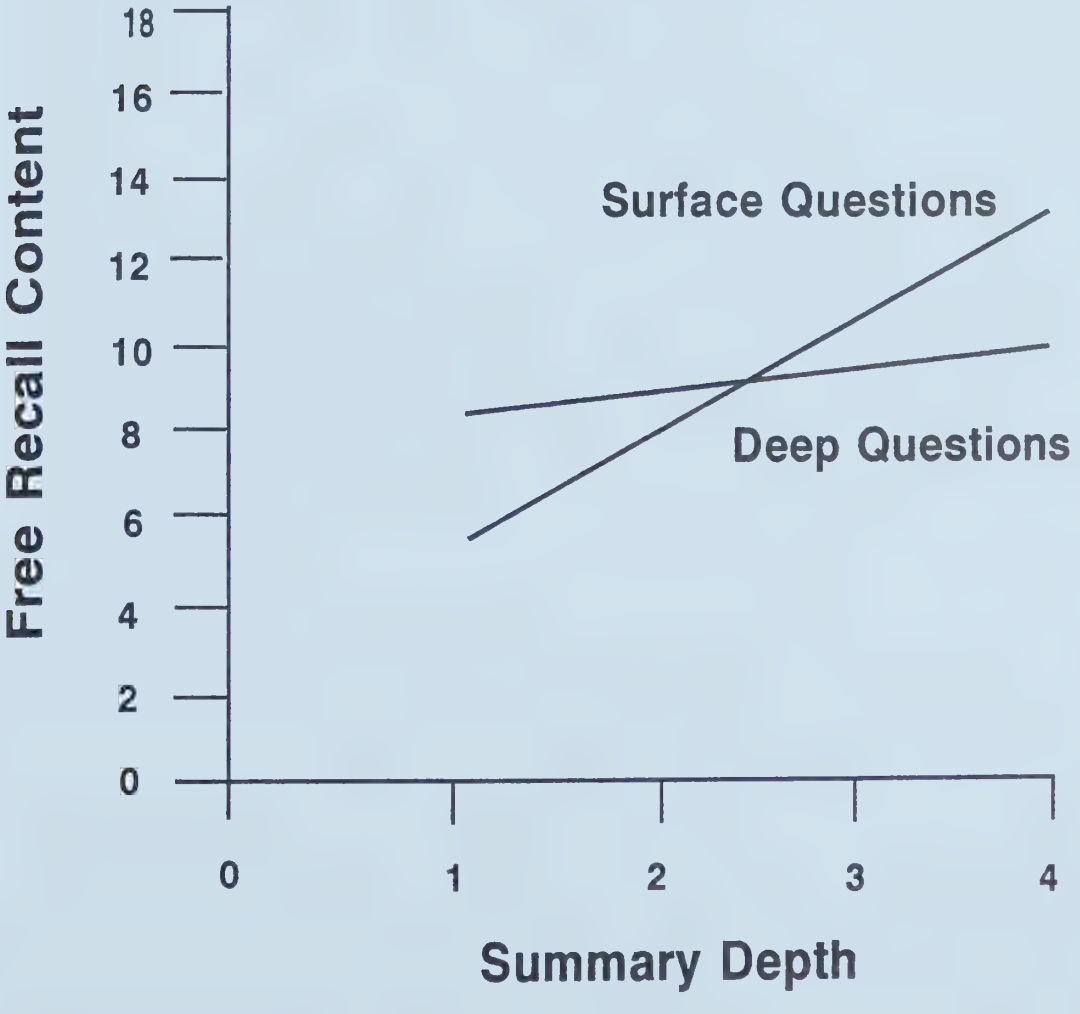


Figure 6. Regression lines showing interaction between task conditions and summary depth (adapted from Fuller, 1992).

Verbalizations during summarization. In the third technique subjects are asked to describe what they are doing as they study and summarize. We developed a 57-category protocol analysis system (Woodhouse & Kirby, 1991), which has subsequently been collapsed into seven or two broader categories (Kirby, Woodhouse, & Ma, 1993). The various categories address different aspects of studying (looking for main ideas, commenting on text difficulty, relating text to prior knowledge, etc.) at different degrees of presumed depth (constructing a main idea as opposed to selecting one from the text). The 57 categories are too cumbersome to employ, but we have had some success with the seven-category version. The two-category version is an attempt to form a surface scale and a deep scale. The seven categories, comprising four surface categories and three deep ones, are shown in Figure 7. The proportion of deep statements made is positively related to performance variables and tends to increase in text absent conditions (e.g., Kirby et al., 1993; Woodhouse & Kirby, 1991).

Depth of Processing During Performance

The final component in Figure 1 concerns learning outcomes. Although we have not studied this explicitly, it is important to remember that subjects exercise options at this point too. For example, a student with a generally deep approach to learning, who has studied the text deeply, may at the point of final performance adopt a more surface or fact-oriented approach. This could be due to performance-related anxiety, to the student's interpretation of the purpose of the test, to fatigue, or to any of a host of other factors that have not been measured. At the moment, all of this is included in the error term. The fact that significant relations between the preceding variables and recall have been shown does not negate the importance of performance factors.

In a sense, most of the study process measures are also measures of performance (a summary is as much a performance as a subsequent free recall). Many of these measures could be applied as successfully to learning outcomes. For example, we also use content scores to assess free recalls; this avoids the difficulty that some subjects copy directly from the text in text present conditions. Biggs and Collis (1982) intended their SOLO Taxonomy as a measure of learning outcomes, not of study processes, which suggests that our depth ratings could also be so used.

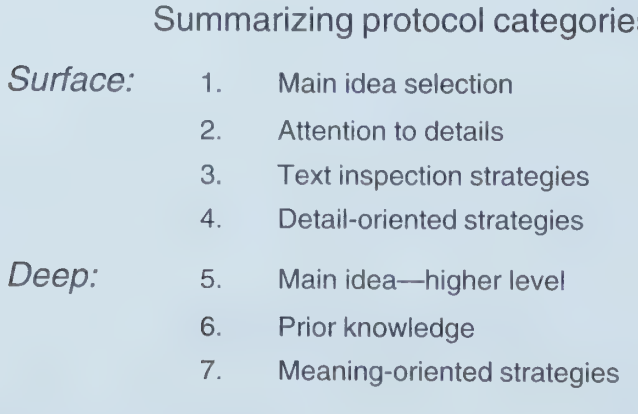


Figure 7. Categories used in protocol analysis.

Implications for Measuring Achievement

How to Do It

It is important to distinguish between what students actually do and what they say they are doing, or even what they say they typically do. What they say they are doing (verbal protocols) may have some value as a research technique, to help validate some measures, but ultimately will be less useful than more objective performance measures. What they say they typically do (their approaches to learning as measured by questionnaire) may have some value in counseling, and may help us understand why some students engage in deeper processing, but it cannot be a substitute for actual performance.

The value of the research we summarize here is that it points to techniques for assessing depth in actual performance, whether during study or on a later test. We see at least five areas of application:

Depth or SOLO ratings. As Biggs and Collis (1982, 1989) suggest, such ratings of quality may be applied to substantial learning products, such as essays or long answers to questions. It is important that the performance be a construction, something generated by the student from memory, rather than something copied from sources or "dumped" from rote memory. Because they attempt to reveal the student's underlying cognitive structure, it is important that the question or prompt not provide too much structure. This condition may be a source of difficulty, because the performance depends critically on the student's interpretation of the goal of the performance. In any particular performance, a student's depth rating could drastically underestimate potential depth. Over time, however, if depth were stressed as a goal, this would become less and less of a problem. It is likely that many students are capable of greater depth than they demonstrate, but do not realize that it is desired (Kirby & Cantwell, 1985; Peters, 1993). Depth or SOLO ratings also offer the potential to provide students with clearer feedback for improving performance.

Content scores. One difficulty with depth ratings is that they are quite global: a five-page essay may yield a single score on a five-point scale. Content scores seem to offer greater potential for guiding students' learning of particular information. Again, it is important that the performance judged be a construction, as copying can yield unrealistically high content scores. There would also be merit in teaching the students text analysis and the content scoring scheme. In this way they could become more aware of text structure and the relative importance of ideas in a text. A text analysis could be used as a concept map to provide students (or to have them generate for themselves) a complete outline of content; these maps could be extended to the thematic level to emphasize that such a level exists and how it is connected to the lower levels.

Short-answer or multiple-choice questions. Although longer, constructed performances are the ideal, it may be useful in some contexts to supplement assessment with shorter, probed recall questions or even multiple-choice questions. Probed recall questions have the disadvantage of providing some structure, so students' performance may overestimate their ability to generate. On the other hand, carefully designed short-answer or multiple-choice questions, based on a text analysis could be helpful. It is important to be clear about what level of information is being assessed (main idea, important supporting idea, unimportant idea) and if an inference is being assessed, what type it is. For

example, inferences may connect ideas at the same level in the text (forming horizontal coherence), or ideas at different levels (forming vertical coherence), or may connect ideas from the text with background knowledge. Most current measures of reading comprehension present a motley collection of items with little regard for level or type.

Dynamic assessment. One difficulty in assessing depth of processing is that some or many students may not know that it is desired and so will not perform as well as they might. One strategy would be to place the students in situations where deeper processing makes more sense or is required. Examples would be text absent summarization and use of poorly structured texts. A second strategy would be to instruct students in the use of deeper processing. By comparing students' performance in standard and depth-encouraging conditions, or pre- and postinstruction, a better measure of ability to engage in deeper processing may be obtained.

Strategy assessment. Underlying all these suggestions is a notion that students' strategies are important, not only in the sense of how they attempt to attain goals, but which goals they seek. Strategies may be inferred from the various depth ratings and content scores, but need to be addressed explicitly in instruction.

Consequences

Although depth of processing is espoused widely as an important goal in education, much of what is done in education works against it. Many students and teachers have conceptions of learning that do not include depth of processing. Many school and classroom practices encourage these conceptions, and defeat attempts for change. At the same time, other notions of teaching and learning are unrealistic, arguing that depth can be attained without content. These notions favor depth, but serve to discredit it as a goal in the eyes of some observers. The effective measurement of depth of processing is an important educational goal, but it needs to develop in conjunction with a greater emphasis on depth of processing in learning and teaching.

Acknowledgments

The research on which this article is based was supported by grants from the Social Sciences and Humanities Research Council of Canada to the first author. The contributions of Jennifer Fuller, Allyson Hadwin, Yamin Ma, Denise Pedwell, Jessica Peters, Jacqueline Simpson, and Bonnie Stein are gratefully acknowledged.

References

Biggs, J.B. (1987). *Student approaches to learning and studying*. Melbourne, Australia: Australian Council for Educational Research.

Biggs, J.B. (1993). What do inventories of students' learning processes really measure? A theoretical review and clarification. *British Journal of Educational Psychology,*

Biggs, J.B., & Collis, K.F.(1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Biggs, J.B., & Collis, K.F. (1989). Towards a model of school-based curriculum development and assessment: Using the SOLO taxonomy. *Australian Journal of Education, 33,* 149-161.

Biggs, J.B., & Kirby, J.R. (1984). Differentiation of learning processes within ability groups. *Educational Psychology, 4,* 21-39.

Biggs, J.B., & Rihn, B. (1984). The effects of intervention on deep and surface approaches to learning. In J.R. Kirby (Ed.), *Cognitive strategies and educational performance*. Orlando, FL: Academic Press.

Craik, F.I.M., & Lockhart, R.S. (1972). Levels of processing: Framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

Duncan, M.J., & Biddle, B.J. (1974). *The study of teaching*. New York: Holt, Rinehart & Winston.

Entwistle, N.J. (1988). Motivational factors in students' approaches to learning. In R.R. Schmeck (Ed.), *Learning strategies and learning styles* (pp. 21-51). New York: Plenum Press.

Fredericksen, C. (1975). Representing logical and semantic structure of knowledge acquired from discourse. *Cognitive Psychology, 7*, 371-458.

Fuller, J. A. (1992). *Deep and surface processing in learning from text among baccalaureate nursing students*. Unpublished master's thesis, Queen's University.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research, 56*, 473-493.

Kintsch, E. (1990). Macroprocesses and microprocesses in the development of summarization skill. *Cognition and Instruction, 7*, 161-195.

Kintsch, W., & van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363-394.

Kirby, J.R. (1988). Style, strategy, and skill in reading. In R.R. Schmeck (Ed.) *Learning styles and learning strategies* (pp. 229-274). New York: Plenum Press.

Kirby, J.R. (1991). Reading to learn: Toward an applied psychology of reading comprehension. In J.B. Biggs (Ed.), *Teaching for learning: The view from cognitive psychology* (pp. 103-125). Melbourne, Australia: Australian Council for Educational Research.

Kirby, J.R., & Cantwell, R.H. (1985). Use of advance organizers to facilitate higher-level text comprehension. *Human Learning, 4*, 159-168.

Kirby, J.R., & Pedwell, D. (1991). Students' approaches to summarization. *Educational Psychology, 11*, 297-307.

Kirby, J.R., Woodhouse, R.A., & Ma, Y. (1993). *Studying in a second language: The experience of Asian students in Canada*. Unpublished manuscript.

Lai, P., & Biggs, J. (1994). Who benefits from mastery learning? *Contemporary Educational Psychology, 19*, 13-23.

Peters, J. (1993). *Summarization strategies, text structure and working memory in recall of text material.* Unpublished master's thesis, Queen's University.

Simpson, J. (1992). *Working memory and the recall of written material*. Unpublished master's thesis, Queen's University.

Stein, B.L., & Kirby, J.R. (1992). The effects of text absent and text present conditions on summarization and recall of text. *Journal of Reading Behavior, 24*, 217-232.

Woodhouse, R.A., & Kirby, J.R. (1991). *A protocol analysis of summarization strategies*. Paper presented to the Canadian Psychological Association, Calgary.

Woodhouse, R.A., Kirby, J.R., Simpson, J., & Hadwin, A.F. (1992, June). *Effects of text structure and summarization processes on recall*. Paper presented at the annual meeting of the Canadian Society for Brain, Behaviour, and Cognitive Science, Quebec City.

The Alberta Journal of Educational Research *Vo. XL, No. 2, June 1994, 163-175*

Philip Nagy
Ontario Institute for Studies in Education

Cognitive Assessment of Solutions to an Ill-structured Problem Using Forced Classification Analysis

This article illustrates the use of forced classification for assessing solutions to ill-structured problems. The analysis, on discussions by children of a social problem, consists of content categorization, quality ratings within content categories, and analysis by categorical discriminant analysis, with student grade and judged quality as discriminating variables. Results show that FCA can be used to identify ways in which respondents to ill-structured problems can be differentiated using detailed differences in content and quality of responses. The technique is applicable to essays, oral examinations, think-aloud analyses, and other ill-structured problems.

Cet article démontre l'utilisation de la classification forcée pour l'évaluation de solutions à des problèmes sociaux mal structurés. En se basant sur des discussions d'élèves, à propos d'un problème social énoncé, l'analyse consiste à la catégorisation du contenu, au classement de qualité à l'intérieur du contenu de ces catégories, et à l'analyse par catégorie de façon à ce que leur contenu soit analysé. On considérait comme variables discriminatoires le niveau scolaire de l'élève et la qualité jugée par l'élève. Les résultats indiquent que l'analyse par la classification forcée (ACF) peut être utilisée pour identifier différentes façons par lesquelles les répondant(e)s aux problèmes sociaux mal structurés peuvent être différencié(e)s en utilisant les différences détaillées du contenu et la qualité des réponses. On peut se servir de cette technique dans des dissertations, dans des examens oraux, des analyses à haute voix, et dans le contexte de d'autres problèmes mal structurés.

Background

The purpose of this article is to illustrate the use of forced classification analysis (FCA, Nishisato, 1984) for the cognitive assessment of solutions to ill-structured problems. The illustration involves 156 discussions by grades 3 and 6 students, taken in groups of about five. The discussions deal with a family situation. The groups of children are here referred to as subjects. This report focuses on the method of analysis that was used, not on the substance of the problems addressed by the subjects.

The development of the methodology is best understood in terms of concepts drawn from assessment, problem solving, memory research, and the quantification of categorical data. The original study from which these data come (Nagy, 1992) draws also on the literature on the teaching of thinking skills (e.g., Resnick & Klopfer, 1989). This latter body of work is not discussed in any detail.

Philip Nagy teaches in the Measurement and Evaluation program. He is interested in cognitive models of assessment.

Much current measurement research can be viewed as an attempt to generalize technologies originally developed for measurement of simple multiple-choice outcomes to more complex outcomes, such as polytomous items and performance assessments. This report comes at the issue from a different perspective. Consider a continuum of outcome complexity, with responses to multiple-choice items at one end (say the left) and responses to complex, ill-structured problems at the other. The left end of this continuum can be considered the traditional realm of measurement theory, whereas the right end has been the arena for one branch of cognitive science. Recent work in measurement can be viewed as moving from the left end toward the middle, whereas this report should be viewed as moving from the right to the middle.

This is an attempt to introduce quantification to descriptive procedures. As Glaser (1991) states,

> The weak link in the construction of proficiency tests was our descriptions of the performances to be measured; systematic techniques were needed to describe more adequately the components of proficiency at specified levels of competence ... In the present state of the art, the study of performance must be given precedence over or, at the least, equal status with measurement technique for effective approaches to subject-matter assessment. (p. 18)

Assessment and Higher-order Objectives

Writing on the relationship between assessment and the curriculum, many authors (e.g., Nagy, Traub, & MacRury, 1986) have observed that what is most easily assessed tends to become viewed as most important. Some of the more complex goals of education lack appropriate assessment methods (Archbald & Newmann, 1988). Consequently, existing assessment aimed primarily at the less complex goals can be viewed as exercising a narrowing influence on the curriculum. Stiggins (1988), for example, has argued that attempts to examine various approaches to the teaching of thinking skills have been limited by available assessment tools. Despite progress in assessing complex goals (e.g., Baker, Freeman, & Clayton, 1991; Collis & Romberg, 1991; Nickerson, 1989), there is tension between testing and curriculum renewal. Although most would agree that the ability to discuss a complex situation is an important higher-order outcome of schooling, it is not easily assessed and, one might surmise, is not often emphasized in the curriculum.

Problem Solving and Schema Theory

The task used in this study is an ill-structured problem. Frederiksen (1984) summarized the distinction between well-structured and ill-structured problems. Ill-structured problems are more complex, have fewer definite criteria for deciding if a solution has been reached, and lack complete information and a convenient list of accepted procedures. They have higher verbal content and are more context dependent. Most real-life problems would be classified as ill-structured, as would many harder-to-define but important outcomes of schooling, such as the ability to reflect on issues, marshall arguments, and discuss.

Methods for the analysis of complex problems are emerging (Larkin, 1980; Lawrence, 1988; Voss, Greene, Post, & Penner, 1983). Voss and Post (1988) have related this work to schema theory (Anderson, Spiro, & Anderson, 1978), a

system for examining memory for complex events in which schemata act as mental structures or generic scripts. Interpretation of memory data involves matching elements in a specific situation to generic slots or placeholders. Schallert (1982) notes that schemata evolve, becoming more elaborate and specific with experience. Schema theory provides procedures useful for charting complex verbal phenomena in a measurement framework; their use does not imply any direct link with memory research. This allows avoidance of the debate on the reality of schemata (e.g., Abelson, 1981; Alba & Hasher, 1983), while taking advantage of a promising procedure for examining individual differences in verbal performance.

Assessing Quality

A difficulty in applying a schema-based analysis in a measurement context is how to introduce conceptions of quality, or how to decide what constitutes a more complete or sophisticated discussion. Horton and Mills (1984) describe this problem as the lack of an independent definition of depth of processing in the schemata perspective. Discussion of quality can be introduced in at least three somewhat unsatisfactory ways: holistic ratings by judges, informed but arbitrary designation of some categories as better than others, or by imposition of an external criterion measure so that data from some respondents (say older or higher achieving) are defined as better.

A fourth method, used here, begins with categorization of elements of a discussion, yielding a subjects-by-categories matrix indicating the presence or absence (1 or 0) of specific elements in each response. Next, a system for categorizing students' approaches to problems into levels of increasing complexity, adapted from the neo-Piagetian perspective of Biggs and Collis (1982), can be applied to rate the statements by each subject within each category. This replaces the 1s in the data matrix with ratings, thereby introducing an assessment of quality that is nested within a specific content category.

Forced Classification Analysis

With a matrix of subjects by content categories, with entries of either 0 or a rating, the issue becomes how to uncover the categories, and rating levels within categories, that best differentiate among subjects. Discriminant analysis (e.g., Kerlinger & Pedhazur, 1973) is inappropriate because of the nature of the data: the data are typically not normally distributed; the matrix is usually relatively empty; and although the ratings have ordinal properties, it is an empirical question whether "good" solutions can be characterized by a statistic such as the frequency or number of highly rated statements. A method that treats the data as nothing more than unordered categories is preferred.

Forced classification analysis (Nishisato, 1984) provides a more appropriate approach. Based on dual scaling (Nishisato, 1980), the principal components analysis of categorical data, FCA allows a variable to be chosen as the forcing variable and maximally discriminates among subjects across this chosen variable.

An introductory example will clarify the approach. Consider the exemplary data in Table 1. Seven subjects have responded to three questions with three options each. These questions might have a correct answer or merely request an opinion. The initial layout of the data appears in Table 1a. The first step in

Table 1a
Data for Seven Subjects Choosing one of Three Options for Three Questions

Questions		*Q.1*			*Q.2*			*Q.3*		
Options:		1	2	3	4	5	6	7	8	9
Subjects	1	0	1	0	1	0	0	0	0	1
	2	1	0	0	0	1	0	1	0	0
	3	0	0	1	0	0	1	0	1	0
	4	1	0	0	0	1	0	0	1	0
	5	0	1	0	0	0	1	0	0	1
	6	0	0	1	0	0	1	1	0	0
	7	0	1	0	1	0	0	0	0	1

Table 1b

Questions		*Q.1*	*Q.2*	*Q.3*
Subjects	1	$Y_1 X_2$	$Y_1 X_4$	$Y_1 X_9$
	2	$Y_2 X_1$	$Y_2 X_5$	$Y_2 X_7$
	3	$Y_3 X_3$	$Y_3 X_6$	$Y_3 X_8$
	4	$Y_4 X_1$	$Y_4 X_5$	$Y_4 X_8$
	5	$Y_5 X_2$	$Y_5 X_6$	$Y_5 X_9$
	6	$Y_6 X_3$	$Y_6 X_6$	$Y_6 X_7$
	7	$Y_7 X_2$	$Y_7 X_4$	$Y_7 X_9$

dual scaling deletes the 0s, and replaces the 1s with an appropriately subscripted x for the option weight multiplied by an appropriately subscripted y for the subject weight. The result appears in Table 1b.

The second step involves the application of Guttman's principle of internal consistency:

1. Assign as similar scores as possible to subjects choosing the same option, and as different scores as possible for those choosing different options; and
2. assign as similar weights as possible to options chosen by one subject, and as different weights as possible to options chosen by different individuals.

Algorithmically, this is operationalized as follows:

1. Determine nine values of X_j, option weights, to minimize within-subjects differences and maximize between-subjects differences in the values of weights; and
2. determine seven values of Y_i, subject weights, to minimize within-option differences and maximize between-option differences in scores.

This amounts to maximizing the product-moment correlation defined over the x, y pairs in the table.

A forced classification analysis is obtained when one of the "items," a categorical variable such as sex, experience, or grade, is designated as the criterion variable. Option weights are derived to maximize correlation of other variables with the criterion. The process assigns weights to subjects and category ratings to maximize the similarity of subjects who use the same categories (at

Table 2
The Data Array for Forced Classification

Grade	*Categories*																
	11	12	211	212	213	214	22	231	232	233	24	25	26	311	312	32	33
3	2	2	0	0	0	2	3	3	2	0	3	0	2	3	0	1	2
3	0	1	3	0	0	2	2	3	0	0	2	0	0	3	1	0	0
3	0	1	2	0	0	0	2	2	0	1	0	0	0	0	0	0	0
3	2	2	1	0	1	2	2	3	0	2	0	0	0	4	0	0	1
3	2	0	2	0	0	1	2	0	0	1	2	0	0	0	0	0	0
3	0	1	2	0	0	2	2	2	0	0	0	0	0	0	0	0	0
3	0	2	1	0	2	2	2	1	0	2	0	0	0	2	3	0	0
3	2	2	1	2	0	0	2	2	0	0	0	0	0	0	1	1	0
6	0	3	0	0	1	0	3	2	0	0	2	2	0	3	0	2	3
6	0	0	3	0	0	3	3	2	0	0	0	0	0	2	0	0	1
6	0	3	3	0	0	2	1	1	2	0	0	0	0	0	1	1	2
6	0	0	2	0	2	3	2	2	2	1	0	0	0	0	0	0	0
6	0	2	1	0	0	3	2	3	3	1	0	0	0	2	0	0	3
6	3	3	0	0	0	0	2	3	0	2	0	1	0	3	3	3	3
6	1	0	0	0	0	2	2	3	0	0	0	0	0	0	0	0	0
6	0	2	0	0	0	2	2	2	3	0	0	0	0	2	4	2	3

the same rating) and to minimize similarity of those who use different categories. Weights are assigned so that those subjects flagged as being maximally different from each other on the FCA solution can be identified and compared. Table 2 shows a data array from the present study, with the forcing variable, Grade, on the left, and column labels from this study, 1.1, 2.1.1, through 3.3, and ratings in the body of the table (decimals are omitted, and column labels are explained below).

In interpreting the results, the categories that best discriminate among subjects can be identified through their correlations with the solution. Also, the specific rating differences within a discriminating category can be determined. For example, one category might discriminate between those who did not address an issue at all and those who did (i.e., between 0 and any other rating), whereas another might distinguish those with high ratings from those with low ratings. Finally, characteristics of subjects at opposite ends of the solution can be compared on independent variables.

A forcing variable with N levels produces N-1 solutions. Successive solutions account for decreasing proportions of variance, but very small drops in variance occur across these solutions, making them essentially equal in potential importance.

There are some disadvantages to FCA. First, the sampling distributions of analyses are unknown, so no statistical tests are possible. The technique serves as a heuristic device only. Second, subjective decisions are required as to the

number of subjects and categories to characterize and interpret in a given solution. And third, information on the ordering of ratings is lost; the data are treated entirely categorically. Despite these problems, FCA classification offers a potentially useful tool for the final stage of these analyses.

Method

The Sample and Data Collection

The analysis proceeds by creating a system of categories that encompasses all of the ideas expressed by the respondents. Statements are categorized into these slots and then rated according to the following scale, adapted from Biggs and Collis (1982):

1. Makes one point without supporting argument or elaboration;
2. Makes two or three unrelated points without supporting arguments or elaboration;
3. Makes more than two points and relates them to each other with elaboration or synthesizes several ideas with evidence of thought;
4. Makes more than two points and relates them to each other with supporting elaboration, showing depth of thought, well-articulated explanations and original ideas.

The subjects came from grade 3 and 6 classes of nine rural elementary schools located in three school districts in southwestern Ontario. In the original study (Nagy, 1992) the three schools in each district were assigned to one of three methods of teaching thinking skills. This treatment variable is not used in the analysis reported here, but its presence and balanced nature should be noted. Data from two consecutive years were used, so that the intended sample involved 18 classrooms from each grade. One class from each grade was lost due to scheduling difficulties.

Data collection. Students from a class were subdivided into groups of about five and asked to discuss, for five or ten minutes, a problem involving two children, B.J. and Pat. Pat is heavily involved in extracurricular activities, while B.J. is around the house a lot. Both get the same allowance, but B.J. is asked to do more chores than Pat. She or he complains to no avail. The problem, presented in writing, was 140 words; age was not specified, and gender needed to be inferred. In total, 156 such discussions were recorded.

Two variables were used as forcing variables in the FCA analysis: Grade refers to either grade 3 or grade 6; Score refers to independent ratings of the overall quality of each group discussion. This latter variable was derived from ratings by two judges, initially on a 10-point scale, but reduced for this analysis to a 3-point scale. Each judge worked independently, and any disagreements were resolved by discussion.

Matrix Generation

The audiotapes were transcribed and segmented with a change in speaker (Ericsson, 1987). Two analysts prepared a hierarchical category system to encompass the responses made to the problem, using a randomly chosen sample of the data. This process required a few iterations, and the category system underwent minor modifications during application to the full data set.

Next, the analysts rearranged the data so that all statements from a subject coded in the same category were placed together, even if uttered at different

times. Then, they rated these collected statements on the 4-point scale. After a two-month break, the analysts gave holistic scores to the discussions on a 10-point scale, under instructions to aim for equal distribution across the scale. For the categorizations, the ratings, and the holistic scores, the two analysts worked independently and then came together to resolve disagreements. The analysis was done originally using a larger number of categories, which were later amalgamated by assigning to each subject the largest rating within the set of categories so combined. The category system appears as Appendix A; 17 (reduced from 31) categories were used.

Forced Classification

At this point, we have a 17-categories-by-156 subjects matrix, with elements of 0 for no mention of a topic through 4 for a well-thought-out argument. For the forcing variable Grade, one solution is reported, and for Score, two solutions. The 4-point depth-of-processing ratings are called *ratings*. Arbitrary decisions are required for interpretation of the FCA results. The first decision is how many subjects to use in exploring solution dimensions. FCA solutions provide weights to subjects, so that those with high positive weights can be contrasted with those with high negative weights. Forty of the 156 groups, 20 at each end of the solution dimension, were contrasted. This contrast involves examining the rating patterns (on the 0-4 scale) for the two groups. It plays a major role in the second decision.

This second decision concerns how many categories to choose to best distinguish between subjects. FCA provides a correlation of each category with the solution, allowing rank-ordering of the categories for both data sets. Reporting starts with the category correlating highest with the solution. With 20 subjects available at each end of the solution, approximate chi-square tests could be used to decide when the patterns of ratings of subjects at each end of the solution were no longer sufficiently different to report.

Results

Agreement between the two analysts on categorization by content was 91% before resolution; corresponding agreement on depth-of-processing ratings was 90%. For the Score variable, the holistic ratings of overall quality, agreement was 90% on the original 10-point scale, with disagreement of more than one point in fewer than 5% of the cases. These 10-point scores were reduced to the 3-point Score variable for this analysis. Because of the bulk of the data, these reliability figures are based on a sample only, and on the original, more complex category systems.

Variance accounted for in the solutions were for Grade, 80% and for Score, 46% for each of two solutions. Coefficient alphas for all solutions were over 0.96.

The responses of the 156 groups of children were categorized using the 17-category system reported in Appendix A. The bulk of the 1,116 non-zero rated statements were in a few categories: discussion of whether the mother's decision was fair or not (9%), changing chore arrangements (26%), activities (13%), or allowances (19%), and discussing personal experiences (13%). A substantive interpretation of the data is available in Nagy (1992).

Table 3
FCA Solution Using Grade as the Forcing Variable Frequencies on Discriminating Categories for Extreme Groups

Category Number	Definition		Rating 0	1	2	3	4
1.1	Mother's decision fair	Gr6	13	1	4	2	0
		Gr3	18	2	0	0	0
2.2	Change activities	Gr6	0	1	10	9	0
		Gr3	12	2	3	2	1
2.3.2	Child should pay for activities	Gr6	9	2	3	5	1
		Gr3	20	0	0	0	0
3.1.2	General knowledge or views	Gr6	2	0	8	4	6
		Gr3	10	8	2	0	0
3.2	Seek additional information	Gr6	6	0	12	2	0
		Gr3	18	2	0	0	0
3.31	Group discussion	Gr6	2	3	1	12	2
		Gr3	11	0	8	1	0

[1]This category is unlike the rest, which are content based. It is an assessment of the extent to which the children talked to each other, rather than merely taking turns talking to the moderator.

The single FCA result using Grade (3 versus 6) as the forcing variable is summarized in Table 3, where frequencies of ratings are reported for contrasting groups at opposite ends of each solution dimension. Six categories were identified as distinguishing between the 20 subjects at opposite ends of the solution dimension. There are clear differences, all favoring the grade 6 students, as would be expected. The solution dimension did not achieve perfect separation by Grade, with three grade 6 groups included in the lower scoring group, and one grade 3 in the higher.

On the first FCA solution using Score as the forcing variable (Table 4), eight categories had significantly different distributions across the ratings. Again, the results are striking. This solution makes a clear distinction between high scoring (19 scores of 3 and 1 score of 2) and low scoring (19 scores of 1 and 1 score of 2) groups. FCA has identified 40 of the 156 groups that differ in consistent ways across eight of the 17 categories. These eight categories include five of six from Table 3, the Grade results, plus three additional ones.

The second FCA solution using Score as the forcing variable is quite different. It makes a slightly less clear distinction between medium scoring discussions (one score of 1, 16 of 2 and 3 of 3) and a mixture of high and low scoring ones (10 scores of 1, 1 of 2, and 9 of 3). Six categories exhibited significantly different ratings patterns, but unlike the first two reported solutions, these patterns are not all in the same direction. The data for three categories, 2.1.1, 2.1.4, and to some extent 2.3.3, favor the medium scoring groups, and Category 2.4 favors the mixture of low and high scoring. The remaining two categories, 2.3.1 and 3.1.2, exhibit unusual patterns contrasting bunched ratings with fairly even distributions across all ratings. This solution seems to have identified unique ways in which middle quality discussions differ from each of high quality and poor quality ones.

Table 4
FCA Solution Using Score as the Forcing Variable
Frequencies on Discriminating Categories for Extreme Groups

Category Number	*Definition*		*Rating* 0	1	2	3	4
First Solution							
2.1.1	BJ do less or Pat do more	HI	5	1	9	4	1
		LO	13	5	2	0	0
2.1.4	Schedule or share chores	HI	2	1	5	8	4
		LO	17	0	3	0	0
2.2	Change activities	HI	1	0	8	11	0
		LO	4	7	9	0	0
2.3.1	Gear privileges to chores	HI	0	0	5	12	3
		LO	10	7	3	0	0
2.3.2	Child should pay for activities	HI	8	1	4	6	1
		LO	19	1	0	0	0
3.1.2	General knowledge or views	HI	0	0	7	5	8
		LO	14	6	0	0	0
3.2	Seek additional information	HI	8	1	9	2	0
		LO	18	2	0	0	0
3.3	Group discussion	HI	3	1	6	8	2
		LO	19	1	0	0	0
Second Solution							
2.1.1	BJ do less or Pat do more	ME	2	4	10	4	0
		HL	11	4	3	1	1
2.1.4	Schedule or share chores	ME	0	2	10	8	0
		HL	13	3	2	2	0
2.3.1	Gear privileges to chores	ME	0	1	19	0	0
		HL	5	5	3	5	2
2.3.3	Keep allowances equal	ME	12	0	7	1	0
		HL	12	4	4	0	0
2.4	BJ talk to friend, family	ME	15	5	0	0	0
		HL	5	2	6	7	0
3.1.2	Role of experience	ME	5	5	10	0	0
		HL	11	2	2	2	3

HI, ME, LO; high, medium or low; HL; a mixture of high and low

Summary and Discussion

The purpose of this article is to explore the capabilities of FCA for the cognitive analysis of ill-structured problems. One major advantage of this procedure is the high degree of reliability attained in the subjective decisions. By separating decisions on content from those on quality, and by constraining the quality decisions, levels of agreement far surpassing those in typical holistic grading (e.g., Nagy, Evans, & Robinson, 1988) were achieved.

Results forcing on Grade showed clear differences. However, the results for the variable Score were of much greater interest, for two reasons: first, the discriminating categories and identified subjects in the first solution were different from those based on Grade, indicating that the two bases for distin-

guishing were not identical; and second, the second solution found patterns of contrasting strength, rather than showing a consistent advantage for one group over the other, as in the Grade results.

It is worth commenting on the numbers of categories used in the analyses. As mentioned above, the analysis began with a much more complex category system of 31 rather than the present 17 categories. Earlier analyses with the larger category systems did not produce such clear results, suggesting that an overly complex category system can make distinctions better left unmade for the purposes of forced classification analysis.

FCA appears useful when used with two- and especially three-level forcing variables. When a three-level variable like Score is used, FCA appears to be very powerful; it can find interesting contrasts and can show different bases for making different distinctions (i.e., medium from high and low) on the forcing variable.

We have relied on methodology taken from cognitive research, and it is reasonable to examine the relationship between this work and schema theory. Because this it group data, any direct connection with memory per se is problematic. Voss and Post (1988) link methodology to theoretical framework. The intent of this work is development of assessment devices for educators. Thus the theoretical concern is to represent the possible approaches to a problem and to find those aspects that reflect differences in quality. When faced with the choice between richness of detail and operational simplification, a concern with operational measurement has favored the latter. Cognitive scientists seeking in-depth meaning in rich detail may object to this adaptation of one of their principal modelling devices. However, from the perspective of a measurement theory attempting to cope with increasing complexity in outcome data it is useful.

Although this work does not advance schema theory, it does advance analysis of ill-structured problems. Lawrence (1988) examined magistrates' thinking, using three magistrates dealing with three cases. Voss et al. (1983) investigated the lack of productivity of the Soviet agricultural system using fewer than 20 subjects. Relying heavily on this earlier work, the present study has moved to considerably larger sample sizes and some quantification while maintaining what appear to be comparable levels of detail and research costs.

How has measurement profited by this exercise? What has been accomplished by this complex analysis that might not have been done more traditionally, by developing a marking scheme or training judges in holistic scoring? There is ample evidence that teachers require assistance in assessing higher-order outcomes of instruction (Haertel, 1986; Stiggins, 1988). Neither the identification of what constitutes higher-order thinking nor the development of appropriate marking systems is a trivial task. Both require a theoretical framework. Further, there are calls from those studying the assessment practice of teachers (e.g., Stiggins & Bridgeford, 1985) for more focused methods. Systematic research on the bases teachers use for grading is thin (Stiggins, Frisbie, & Griswold, 1989). A method allowing comparison of practice with a well-developed image of what constitutes the target qualifies as a focused method, both for research and for the improvement of practice.

The analysis has proven useful in dissecting complexity in a systematic manner. It has provided a "territorial map" of a complex domain that might be expanded by use of similar data from responses to carefully constructed scenarios. Such work is made easier and less expensive through the use of such a tool. Beyond this particular data set, the technique provides a general method for examining ill-structured problem solving, which plays a prominent role in many curricula, particularly the social sciences.

On the side of caution, forced classification lays open possible structure in the data, but does not replace careful exploration by other means. As with any matrix decomposition method, peculiarities in the data can lead to uninterpretable results. The final test of forced clasification is whether the results make sense in terms of the raw data.

The need for additional careful examination does not diminish the value of the analysis. The method has done more than find two unusual sets of subjects, along with an idiosyncratic way in which they differ; it has identified and exaggerated real differences that exist in the entire sample. If cross-tabulations of grade by rating level are examined for all 17 categories in the data (a tedious process), it is possible to validate the results of Table 3. The cross-tabulations for all 156 groups mimic the data for the 40 groups in Table 3, although the proportions are somewhat more attenuated. Forced classification is efficient in finding differences, but benefits beyond mere time saving are evident when the forcing variable has more than one solution, such as Score. Again, the full set of 17 cross-tabulations reveals no evidence disconfirming the patterns reported in Table 4. However, although these results may be validated by examination of cross-tabulations, they could never be discovered by such examination.

The ability to discuss an ill-structured problem is one example of an important class of somewhat neglected outcomes of schooling. Other members of this class include essays or oral presentations where an opinion must be defended. How "ill" the structure is varies: we can defend marking schemes for explanations of the causes of historical events better than we can views on current ethical or social dilemmas. This report is intended to spark interest in a meeting of cognitive and measurement perspectives on an important issue.

Acknowledgment

The study on which this report is based was funded by the Ontario Ministry of Education.

References

Abelson, R.P. (1981). Psychological status of the script concept. *American Psychologist, 36,* 715-729.

Alba, J.W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin, 93,* 203-231.

Anderson, R.C., Spiro, R.J., & Anderson, M.C. (1978). Schemata as scaffolding for the representation of information in connected discourse. *American Educational Research Journal, 15,* 433-439.

Archbald, D.A., & Newmann, F.M. (1988) *Beyond standardized testing*. Reston VA: National Association of Secondary School Principals.

Baker, E.L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing.In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition* (pp. 131-153). Toronto: Prentice-Hall.

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning*. New York: Academic Press.

Collis, K., & Romberg, T.A. (1991). Assessment of mathematical performance: An analysis of open-ended test items. In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition*. Toronto: Prentice-Hall.

Ericsson, K.A. (1987). Theoretical implications from protocol analysis on testing and measurement. In R.R. Ronning, J.C. Conoley, J.A. Glover, & J.C. Witt (Eds.), *The influence of cognitive psychology on testing* (pp. 191-226). Hillsdale, NJ: Erlbaum.

Frederiksen, N. (1984). Implications of cognitive theory for instruction in problem solving. *Review of Educational Research, 54*, 363-407.

Glaser, R. (1991). Expertise and assessment. In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition* (pp. 17-20). Toronto: Prentice-Hall.

Haertel, E. (1986). *Choosing and using classroom tests: Teachers' perspectives on assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Horton, D.L., & Mills, C.B. (1984). Human learning and memory. *Annual Review of Psychology, 35*, 361-394.

Kerlinger, F.N. & Pedhazur, E.J. (1973). *Multiple regression in behavioral research.* Toronto: Holt, Rinehart & Winston.

Larkin, J.H. (1980). Teaching problem solving in physics: The psychological laboratory and the practical classroom. In D.T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research* (pp. 111-125). Hillsdale, NJ: Erlbaum.

Lawrence, J.A. (1988). Expertise on the bench: Modelling magistrates' judicial decision-making. In M.T.H. Chi, R. Glaser, & M.J. Farr (Eds.), *The nature of expertise* (pp. 229-259). Hillsdale, NJ: Erlbaum.

Nagy, P. (1992). The assessment of group discussions and complex problem solving: Potential contributions of schema theory. In D.J. Bateson (Ed.), *Classroom testing in Canada: Proceedings of the second Canadian conference on classroom testing* (pp. 20-37). Vancouver, BC: University of British Columbia, Centre for Applied Studies in Evaluation.

Nagy, P., Evans, P., & Robinson, F. (1988). Exploratory analysis of disagreement among holistic essay scorers. *Alberta Journal of Educational Research, 34*, 355-374.

Nagy, P., Traub R., & MacRury, K. (1986). *Strategies for evaluating the impact of province-wide testing.* Toronto: Ministry of Education.

Nickerson, R.S. (1989). *Educational Researcher, 18*(9) (Special Issue).

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications.* Toronto: University of Toronto Press.

Nishisato, S. (1984). Forced classification: A simple application of a quantification method. *Psychometrika, 49*, 25-36.

Resnick, L., & Klopfer, L. (Eds.). (1989). *Toward the thinking curriculum: Current cognitive research. ASCD yearbook.* Alexandria, VA: Association for Supervision and Curriculum Development.

Schallert, D.L. (1982). The significance of knowledge: A synthesis of research related to schema theory. In W. Otto & S. White (Eds.), *Reading expository material* (pp. 13-48). Toronto: Academic Press.

Stiggins, R.J. (1988). *The nature and quality of teacher-developed classroom assessments.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Stiggins, R.J., & Bridgeford, N.J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*(4), 271-286.

Stiggins, R.J., Frisbie, D.A., & Griswold, P.A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement Issues and Practice, 8*(2), 5-14.

Voss, J.F., Greene, T.R., Post, T.A., & Penner, B.C. (1983). Problem solving skill in the social sciences. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research theory* (pp. 165-213). New York: Academic Press.

Voss, J.F., & Post, T.A. (1988). On the solving of ill-structured problems. In M.T.H. Chi, R. Glaser, & M.J. Farr (Eds.), *The nature of expertise* (pp. 261-285). Hillsdale, NJ: Erlbaum.

Appendix A
Category System for Group Responses

Category Number	*Definition*
1.	Diagnosis
1.1	Mother's decision fair or neutral
1.2	Mother's decision unfair
2.	Solution
2.1	Change chores
2.1.1	BJ do less or Pat do more
2.1.2	Leave as is
2.1.3	Get additional help
2.1.4	Schedule or share chores, co-operate
2.2	Change activities
2.3	Change allowance
2.3.1	Gear allowance, privileges to chores
2.3.2	Children should pay for activities
2.3.3	Stay the same, both increase, decrease
2.4	Communicate: BJ talk to friend, family
2.5	BJ get a job
2.6	Negative behaviour
3.	Problem solving
3.1	Reflections of experience/knowledge
3.1.1	Specific personal experience
3.1.2	General knowledge or views
3.2	Seek additional information
3.3	Group discussion

The Alberta Journal of Educational Research *Vol. XL, No. 2, June 1994, 177-193*

Philip H. Winne

Lorene Gupta

and

John C. Nesbit
Simon Fraser University

Exploring Individual Differences in Studying Strategies Using Graph Theoretic Statistics

Computer-based adaptive learning systems that afford a high degree of learner control in studying are an excellent medium for collecting extensive data on individual differences. STUDY (Winne, Jones, & Field, 1992) is such a system for writing and delivering adaptive tutorials. As learners engage in a tutorial with STUDY, navigating through content and applying studying actions such as underlining, taking notes, requesting elaborations, and attempting test items, STUDY creates detailed time-stamped records of the learner's interactions in a log file. We are developing a methodology based on the mathematics of directed graphs for analyzing log file data. In this approach a sequence of study actions is reduced to a set of nodes representing action types and a set of links representing a temporal relation called followed-by. *This article describes elementary graph theory statistics and features of graphs that can characterize individual differences in cognitive strategies used in studying.*

Les systèmes informatisés d'apprentissage adaptif qui permettent un degré élevé de contrôle sur l'enseignement personnalisé en fonction de l'apprenant(e) sont un moyen excellent de receuillir plusieurs données sur les différences individuelles. STUDY (Winnie, Jones & Field, 1992) est un tel système d'ordinateur qui permet la composition et la présentation de modules d'études adaptifs sur ordinateur. Quand les apprenant(e)s s'engagent à travailler avec les modules de STUDY, ils doivent naviguer à travers le contenu et faire l'application de différentes méthodes d'étude telles que souligner, prendre des notes, solliciter des élaborations, et essayer de répondre aux questions d'examen. STUDY enregistre de façon chronométrée à l'aide d'un totalisateur intégrée à même le logiciel les interactions des apprenant(e)s dans un fichier de receuils de données. Nous sommes en train de développer une méthodologie pour l'étude de ces receuils de données selon la signification mathématique contenu à l'intérieur des graphiques indiqués par le système dans les modules. Dans cette approche, une série de différentes méthodes d'étude peut être réduite à un groupe de modules qui représente différents types de moyens ainsi qu'un groupe de liens qui représente la relation temporelle que l'on appelle suivi-par. Cet article décrit comment la théorie des statistiques des graphiques élémentaires et les différents aspects de ces graphiques peuvent bien caracté-

---

Philip Winne is a professor of education and psychology and an instructional psychologist specializing in theories of complex academic learning and adaptive learning systems that promote learning and motivation.
Lorene Gupta is a doctoral candidate in the Joint MA/PhD program in psychology and education. She is studying applications of computer technologies that promote students' active cognitive engagement in learning.
John Nesbit received his doctorate from the University of Alberta in 1988 and is presently a research associate in Project STUDY. His research focuses on the psychology of computer-based instruction.

riser les différences individuelles dans le contexte des stratégies cognitives qu'utilisent les apprenant(e)s lorsqu'ils ou lorsqu'elles étudient.

As students study, they engage in a stream of varied and interacting cognitive activities: selecting information, seeking information, rehearsing, elaborating, self-testing, and so forth. Some of these cognitive events can be traced (Winne, 1982, 1992) by observable actions the learner performs during studying such as underlining, looking up a definition, copying information into a notebook, circling a feature in a diagram and drawing a line to link it to a portion of a text, and attempting to answer a test item. Interspersed among these cognitive and behavioral events is metacognition, strategies the learner engages to guide selecting and applying each studying action, and adapting tactical sets of those acts. At yet another level, there is a dynamic interplay among (a) these cognitive events; (b) more persistent individual difference variables such as prior knowledge and perceptions of self-efficacy; and (c) instructional interventions such as adjunct questions and feedback (Corno & Snow, 1986; Winne, 1993a, 1993b). At this level, questions address how students self regulate (Zimmerman & Schunk, 1989).

A methodology that could selectively examine a stream of observed activities at theoretically different levels—acts, tactics, strategies, and self-regulation—could contribute significantly to advancing instructional theory and educational practice. Contemporary empirical methods suffer shortcomings in representing temporally unfolding cognitive engagements that constitute a learner's expressions of knowledge, motivation, cognition, metacognition, and self-regulation. Classical Fisherian experimental designs, for example, aggregate data that reflect theoretically separate cognitive events constitutive of individual differences. As well, this methodological approach does not examine patterns in time that reflect using a strategy. Retrospective questionnaires and stimulated recall protocols about cognition engaged in an earlier setting (Eriksson & Simon, 1993) have potential to represent such patterns, but because they do not gather data as cognition happens, there is the possibility that learners misrepresent cognition and motivation by engaging in reconstructive processes. Think-aloud protocols (Eriksson & Simon, 1993) may eliminate misrepresentations attributable to reconstructive recall, but they risk interfering with the very individual differences a researcher seeks to study (Genest & Turk, 1981).

We are attempting to develop methodologies for research into instruction and individual differences that reduce these shortcomings and, particularly, that reveal patterns of cognition latent in sequential records of a learner's actions while studying. Our instrument for gathering data, briefly described in the next section, is a computer-based adaptive learning environment called STUDY (Winne, Jones, Field, & Nesbit, n.d.). STUDY is a milieu where a student studies and in the course of that activity generates and traces cognitive and metacognitive events. Subsequently, we introduce analytical techniques we are exploring for researching cognitive and metacognitive individual differences in students' studying behavior.

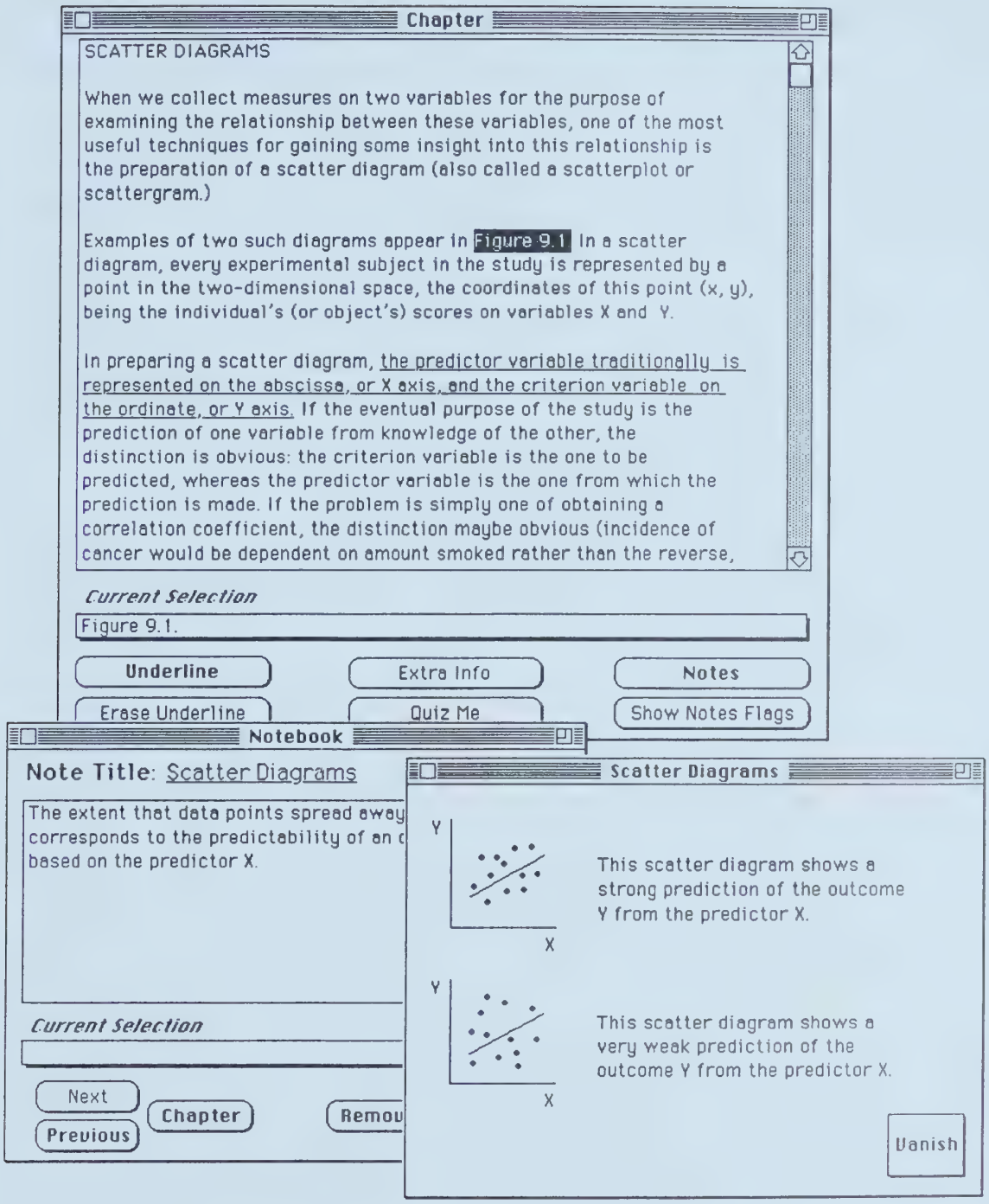

Chapter
SCATTER DIAGRAMS
When we collect measures on two variables for the purpose of examining the relationship between these variables, one of the most useful techniques for gaining some insight into this relationship is the preparation of a scatter diagram (also called a scatterplot or scattergram.)
Examples of two such diagrams appear in Figure 9.1 In a scatter diagram, every experimental subject in the study is represented by a point in the two-dimensional space, the coordinates of this point (x, y), being the individual's (or object's) scores on variables X and Y.
In preparing a scatter diagram, the predictor variable traditionally is represented on the abscissa, or X axis, and the criterion variable on the ordinate, or Y axis. If the eventual purpose of the study is the prediction of one variable from knowledge of the other, the distinction is obvious: the criterion variable is the one to be predicted, whereas the predictor variable is the one from which the prediction is made. If the problem is simply one of obtaining a correlation coefficient, the distinction maybe obvious (incidence of cancer would be dependent on amount smoked rather than the reverse,
Current Selection
Figure 9.1.
Underline
Extra Info
Notes
Erase Underline
Quiz Me
Show Notes Flags
Notebook
Note Title: Scatter Diagrams
The extent that data points spread away
corresponds to the predictability of an
based on the predictor X.
Current Selection
Next
Previous
Chapter
Remou
Scatter Diagrams
Y
X
This scatter diagram shows a strong prediction of the outcome Y from the predictor X.
Y
X
This scatter diagram shows a very weak prediction of the outcome Y from the predictor X.
Vanish

Figure 1. Screen shot of STUDY at an early point in a tutorial.

STUDY: A Tool for Instructional Research into Cognitive Tactics and Strategies

STUDY is a computer system for writing and delivering adaptive learning tutorials on Macintosh™ computers. The system is an amalgam of HyperCard™, which provides the user interface, Smart Elements™, an object-oriented expert system used to represent information and to control the behavior of the tutorial, and routines written in C that glue together the two subsystems and support authoring operations. In STUDY's authoring environment, a designer creates a tutorial by specifying windows, buttons, and fields of text (see Figure 1) and writing rules that define how STUDY interacts with the learner. In a typical tutorial a student would study by viewing text and graphics, clicking on text fields and buttons to display windows containing

ancillary information, attempting test items, and receiving feedback. One goal of the program of research of which STUDY is a part is to investigate individual differences in cognitive strategies that learners use, develop, and can be taught.

STUDY is an adaptive system in that during a tutorial data representing a learner's behavior can be operated on by rules that tailor instructional events to the learner's prior and evolving knowledge, motivation, and strategies. For instance, the designer may identify a continuous series of characters in a text as a concept and program a rule so these words or phrases function as hypertext buttons. During a tutorial, if a student clicks on a concept, a rule may fire that displays a window asking the student to select a type of information (e.g., definition, example, linked concepts) to be displayed by STUDY. The rule also can contain conditions so that another student with a different log of data who clicked on that concept would be shunted directly to an introductory explanation and an accompanying test item about prerequisites.

Following is a partial inventory of learner actions (see Winne, 1992; Winne, Jones, & Field, 1992) used later in this article, which have been researched in instructional psychology: underlining (Peterson, 1992), notetaking (Kierwa et al., 1991), feedback (Butler, Winne, & McGinn, n.d.), viewing examples (Walczyk & Hall, 1989), testing (Glover, 1989), and making attributions (Graham, 1991). During a tutorial, STUDY writes to a log file time-stamped records of all student actions (excluding window scrolling) and all significant actions performed by STUDY (e.g., firing rules, opening windows). By carefully designing tools for the learner to use during a tutorial so that their use validly represents cognitive activities of theoretical interest (Winne, 1982; 1993a), data in the log trace how the learner studies, what the learner achieves, and the conditions on which the learner exercises tactical, strategic, and self-regulatory actions.

Learner Action	*STUDY or Learner Behavior*
Select Text	A portion of text being studied may be selected by a click-and-drag operation. This action is usually a prerequisite for using the text in a subsequent action.
Underline Text	Selected text is underlined. The underline is maintained whenever the text is later viewed by the learner.
Classify Text	The learner or STUDY classifies selected text, for example, as a main idea, a detail, an example or elaboration, or a definition.
Make Note	A window opens in which the learner can enter text. Information entered in a note window is pinned to text the learner has selected in the chapter text or, if no text is selected, to the chapter as a whole. STUDY asks the learner to classify the note as a question posed, a paraphrase of the selected text, a summary, or links made between information.
Request Elaboration	A window opens to present information related to the selected text. The elaboration may be in the form of a

	definition, an analogy, a diagram, an explanation, or a review.
Request Test	An item testing current material is administered. The learner may specify whether the item tests recall or inference.
Request Feedback	After responding to a test item, the learner may choose to receive no feedback, right/wrong feedback, or explanation feedback.
Attribute Performance	The learner specifies an attribution to ability, effort, luck, ease/difficulty of item, or use of a studying strategy.
Calibrate Understanding	The learner estimates comprehension of currently selected sections of text.

An Introduction to Analyzing STUDY's Log Files

In our terms a log file is a chronological record of events occurring in one tutorial with one student. In processing a log file, the analyst filters out items that have no relevance to a particular analysis, rewrites the items of interest to a new file, classifies each as belonging to one of several action types, and denotes each action type by a unique symbol. The resulting temporally ordered string of symbols is called an action sequence. The grand action sequence may be subdivided into smaller action sequences if desired. Action sequences are the primary units of analysis used in describing a student's studying activities.

The most direct measures for describing an action sequence are based on counts of actions such as the number of times a learner retreated in studying a text (e.g., Goldman & Saul, 1990) or the proportion of studied idea units underlined (Johnson, 1988). Useful though such descriptions may be in describing individual differences in the prevalence of comprehension monitoring, counts of action types and statistics derived from them (proportions, means, variances, and so forth) ignore temporal and other relations among multiple actions. We strive to capitalize on such temporal relations. Our hypotheses are that temporally unfolding actions create tactical and strategic structures, and that such structures may be inter- and intraindividual differences that add important information to theoretical accounts about how achievements are created in the course of studying (Winne, 1993a; Winne & Butler, 1994).

In this article we explore analytical tools, based on graph theory, that focus on transitions between actions. In this regard our approach to studying learner-system interactions parallels recent and independent work by Guzdial (1993) who has used Markov chains to analyze programming strategies adopted by science students. Following, we introduce descriptors or *statistics* in elementary directed graph theory and consider how these statistics might reflect individual differences of interest to a cognitive theorist.

An Adjacency Relation and Individual Differences

Consider transitions between two actions, *a* and *b*, observed within a longer sequence of actions generated in a tutorial with STUDY. Suppose *a* is underlining and *b* is making a note that relates *a*'s information to the student's prior knowledge. Transitions among events constitute a relation we call *followed-by*.

If *a* is always followed by *b*, and *b* is always preceded by *a*, the student has composed (Anderson, 1983) a two-action tactic, $a \rightarrow b$, for studying. If we assume the student is a goal-directed, active cognitive agent (Winne & Butler, 1994), each tactic serves a particular goal. Different goals are approached by different actions, tactics, or strategies. Observing a tactic such as $a \rightarrow b$ marks several kinds of information about individual differences. First, the student (a) has conditional knowledge about states of cognitive engagement during studying that can trigger the tactic $a \rightarrow b$. On observing multiple uses of the $a \rightarrow b$ tactic within a longer sequence of other actions, we also infer that the student (b) attended to cognitive states, (c) discriminated some cognitive states from others, and (d) adapted behavior to accommodate particular states.

We hold that conditional knowledge, attention, discriminated action (adaptation), and the goals that learners seek are core elements that constitute individual differences. If a second student frequently uses *a* separately from *b* or replaces *b* with a sequence $c \rightarrow d \rightarrow e$, this variance in cognitive engagement reflects an individual difference we are interested in studying. Depending on other information about the students and their success at the common task they were assigned, the first student's composition of the three-step sequence of procedural knowledge $c \rightarrow d \rightarrow e$ into the unit *b*, and that student's automatization of $a \rightarrow$b may characterize expertise. Or it may reflect a hobbling mechanization of thought reminiscent of classic findings about negative transfer in problem solving (Luchins & Luchins, 1959). Parallel reasoning applied to one student's intraindividual variation across situations would view such variance as reflecting cognitive development.

Actions

Creating an adjacency matrix is the first step in an analysis using the followed-by relation. Here is a simple example using a small subset of actions available in STUDY and a short action sequence:

E = Ask for an elaboration of selected text.
N = Make a note.
U = Underline a portion of text.
Sequence A: U U N U U U E N U N

Construct an N x N matrix where N is the number of possible actions (three in this example). Label rows and columns with the action symbols. Here we arbitrarily use alphabetical order: E, N, U. In cell (i, j) tally all occasions when action i is followed by action j immediately. Thus the first tally representing the first transition U→U is tallied in cell (3, 3). The next tally, representing the transition from the second U to the first N, is tallied in cell (3, 2). The result,

Table 1
A Weighted Adjacency Matrix and an Adjacency Matrix

	E	N	U		E	N	U
E	0	1	0	E	0	1	0
N	0	0	2	N	0	0	1
U	1	2	3	U	1	1	1

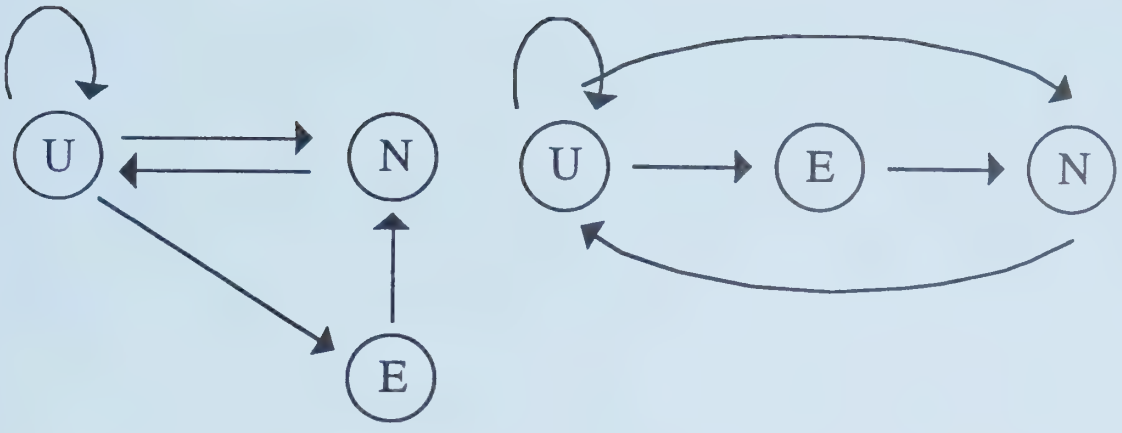


Figure 2. Graph of sequence A showing how repositioning nodes alters the impression conveyed.

illustrated in Table 1, is a weighted adjacency matrix showing the frequencies of all action pairs observed in Sequence A. A weighted adjacency matrix can be simplified to show merely whether action i was followed by action j at least once in the sequence by replacing each non-zero entry in the weighted adjacency matrix with 1, producing an *adjacency matrix*.

Graphing the Followed-by Relation

A graph represents information in terms of a finite set of elements and relationships among the elements. In the example, the elements are actions E, N, U, and there is one relation, followed-by. The left panel of Figure 2 graphically represents the adjacency matrix in Table 1. In Figure 2, each action is represented as a labeled node. For each 1 in row i, column j of the adjacency matrix, a directional link is drawn from the node representing action i to the node representing action j. A self-loop is shown for action U. The graph is *directed* because links represent the directional relation, followed-by. A graph, like Figure 2, is *connected* if none of its nodes is isolated from other nodes. A *weighted graph* labels links using weights in the weighted adjacency matrix.

A directed graph can be visually deceptive because an adjacency matrix implies nothing about where to position nodes. Alternate spatial arrangements may convey different impressions. For example, the right panel of Figure 2 suggests a more "linear" approach to studying.

A Simulated Data Log

To illustrate graph theoretic analyses of action sequences, we generated a log representative of a real STUDY tutorial.[1] One of us (Gupta) kept a record of studying actions while reading the first five sections of Chapter 9 in *Statistical Methods for Psychology* (Howell, 1987) in a course taught by another of us (Winne). Gupta used the three actions (U, E, N) already illustrated, and two more. Although a book can not offer test questions and feedback, as STUDY can, Gupta self-generated test questions (T) and searched the text for feedback about her answers to her self-generated questions (F). T→F is a metacognitive tactic she uses to enhance comprehension (Haller, Child, & Walberg, 1988).

Figure 3 displays the weighted graph, which we call the ALL graph, representing the action sequence across all five sections of the text. In the ALL graph, we excluded the four followed-by links between actions that span sections of the chapter. This reflects our assumption that sections are major

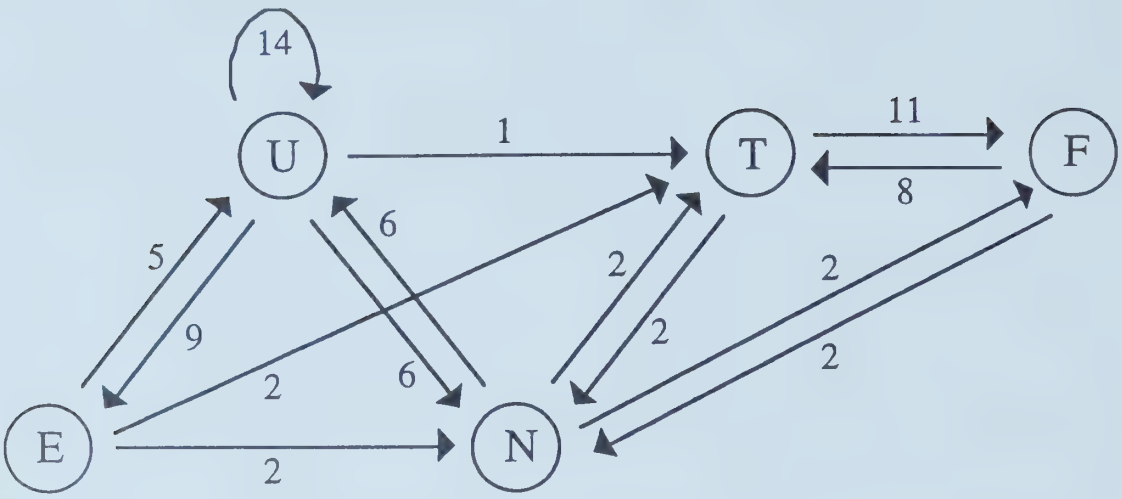


Figure 3. ALL graph of studying actions (transitions between sections are excluded).

content boundaries that probably bound studying strategies and tactics as well. The data corroborate our assumption: the transition F→U never occurred within a section but was observed at most section boundaries.

Graph Concepts and Statistics for Making Interpretations about Strategic Learning

Is a Student Strategic?

Contemporary instructional theories assume that students engaging in learning are goal-directed agents. If this is true, patterns in log files of students' studying should be found that could be interpreted as tracing preferences for or conditional knowledge about actions, tactics, and strategies. For example, in an interview after she studied the text, Gupta indicated her presumption that access to feedback was in one-to-one correspondence to a specific self-generated test question, a pattern like the rigid $a \rightarrow b$ tactic we used to introduce the followed-by relation. Empirically, Gupta was not as consistent as she believed herself to be. The ALL graph shows that she twice made a note preceding feedback and twice made a note following a (self-generated) test item. In these instances, we can infer that she adapted her general tactic to specific, local cognitive conditions.

Gupta's studying operationalized other patterns that reduced the theoretical maximum of N^2 pairwise links within the 5-tuple {U, N, E, T, F}. After examining the log file and interviewing Gupta about her studying and her understanding of STUDY's design capabilities, it was clear that she deliberately excluded links T→U and T→E, as well as self-loops E→E and N→N. Gupta did not know that STUDY could be designed to permit students to return to the text or ask for elaboration immediately after a test item is posed. Nor did she realize that STUDY can be designed to allow students to make two successive but unique elaborations or two successive but unique notes. Gupta also presumed that STUDY would never provide two successive but unique feedback messages, F→F, say, one about content and another about motivation. These tacit rules express individual differences. They reduced the number of theoretically possible followed-by links from 25 to 20, constraining Gupta's ALL graph to a maximum of 80% of STUDY's canonical graph.[2]

The degree to which constraint has been imposed on a canonical adjacency matrix can be measured by two statistics adapted from graph theory, density and cohesion. We interpret these statistics to reflect preference or conditional knowledge.

Density is a graph theoretic statistic, ranging between 0 and 1, that compares the number of links observed in a graph to the number of possible links if all nodes were mutually linked. For N nodes, the number of possible links, including self links (loops), is N^2.

$$\text{density} = \frac{\text{\# links in graph}}{N^2}, \qquad (1)$$

where N is the number of nodes in the graph.

Based on Gupta's reduced canonical graph, we replaced N^2 in equation 1 with 20 instead of 25. With this change, the density of Gupta's ALL graph is .70 relative to her canonical graph for the 5-tuple {U, N, E, T, F}. By reflecting density (reversing its scale by subtracting observed density from the maximum of 1.00), we propose a statistic K_d that measures the facet of individual differences we referred to earlier as discriminated action. K_d is .30 for Gupta's ALL graph relative to the canonical graph. Her cognitive activities, reflected through studying actions that trace her cognition, were constrained by her rules. We might ask the same question of Gupta's studying each of the text's five sections. Table 2 presents K_d values for each section's graph of studying normalized relative to the canonical ALL graph. When observed sequences have sufficient length (sample size), variations in density across sections indicate intraindividual differences in discriminated action when studying different sections of a text.

A second measure of discriminated action that a student exercises during studying is given by the graph theoretic statistic called cohesion. Cohesion measures the proportion of links in a graph, excluding self-loops, that are two-way; that is, for nodes A and B, there is a link from A to B and a second link from B to A. Cohesion is defined in equation 2.

$$\text{cohesion} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij} \cdot x_{ji}}{\frac{N^2 - N}{2}}, \quad i \neq j, \qquad (2)$$

Table 2
K_d and K_c Constraint Statistics for Gupta's Log

Section	*9.1*	*9.2*	*9.3*	*9.4*	*9.5*	*ALL*
K_d	.500	.850	.650	.550	.500	.300
K_c	.625	.875	.875	.625	.500	.375

Note. The ALL graph is normalized relative to STUDY's theoretically maximal (canonical) graph. Graphs representing individual sections are normalized relative to the canonical ALL graph.

where x_{ij} is the entry in the i^{th} row j^{th} column of an adjacency matrix and $\frac{N^2-N}{2}$ is the number of possible pairs of nodes. Like density, cohesion ranges from a minimum of 0 to a maximum of 1. The more cohesive a graph, the less constraint a learner imposes on selections from the set of actions. By reflecting cohesion, we define a second measure of discriminated action, K_c. The value of K_c for Gupta's ALL graph is .50 normalized relative to STUDY's theoretically maximally cohesive graph. Gupta's rules disallowed two two-way relations (T→U and T→E) of 10 possible. Table 2 presents values of K_c, normalized relative to Gupta's set of rules, for each of the five sections of text.

By carefully defining a canonical graph relative to which empirically generated graphs are normalized, we suggest it is possible to characterize degrees to which a student exercises discriminated action in studying. Such discriminations depend on attention and conditional knowledge (Winne & Butler, 1994), that is, perceptions about the match between information in the instructional environment at a given time and the student's cognitive states. Based on the poststudying interviews with Gupta, we speculate that K_d may reflect the richness of a student's conditional knowledge about studying tactics to use while learning relative to the student's understanding of the task assigned.

Identifying Cognitive Tactics

Treatments in instructional research are designed to engage students in particular patterns of cognition triggered by particular conditions. For example, to promote the hypothetical cognitive activities of (a) focusing attention by using a rule to monitor information, (b) rehearsing important information, and (c) enhancing retrievability of that information by meaningfully elaborating it, students might be trained to (a) discriminate important information in a text, (b) underline it, and (c) make a note about it. These events comprise a cognitive tactic, a pattern among actions that a learner uses to achieve a relatively specific goal. Tactics can be represented as condition-action rules: if condition A is true (information is judged important to learn), then perform action X (reread that information as it is underlined); if condition B is true (rehearsal is

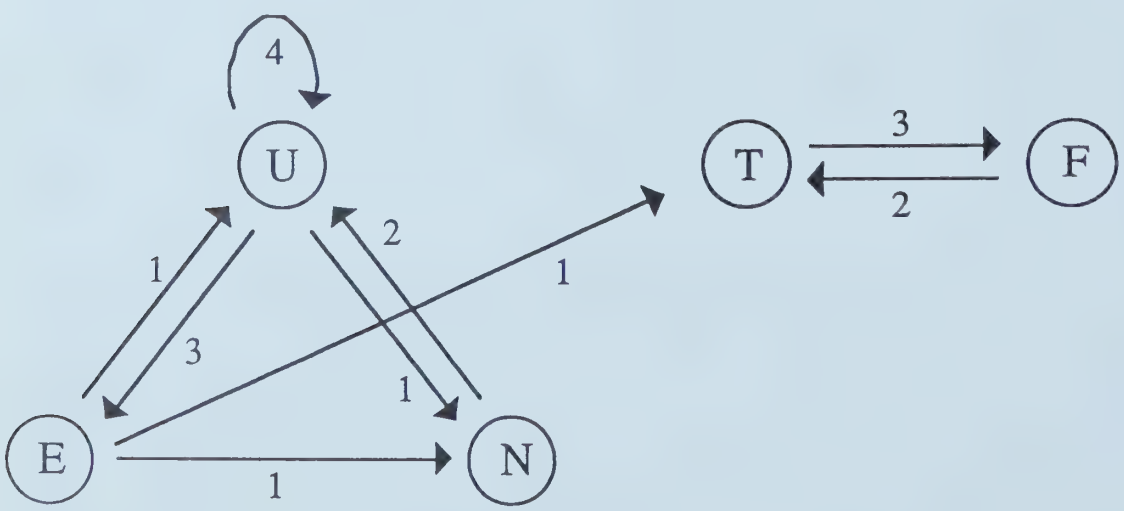


Figure 4. Graph of studying in Section 9.4 {U U U E N U N U E U U U E T F T F T F}.

complete and retrievability should be enhanced), then perform action Y (make information meaningful by paraphrasing it in a note). By equipping STUDY with tools for a student to trace each element in this cognitive tactic, a graph of those traces can represent the pattern of actions whenever the student applies that tactic. Graph theory provides tools that help to identify tactics within action sequences and that describe the degree to which tactics resemble one another or resemble a canonical pattern.

One defining feature of a cognitive tactic is that it is engaged only under particular conditions. This feature can be manifested in a graph of studying events by a node that is a gateway to a tactic graphed as a small, tightly linked pattern among other nodes. In the graph of Gupta's studying in Section 9.4 of the text (Figure 4), node E plays this role. A request for an elaboration led Gupta to engage in a reverberating cycle of asking for a test (T) question and requesting feedback about her answer (F).

In graph theory a *bridge* is a link in a connected graph that, if removed, creates two separate connected graphs. Bridges can represent links where the preceding act triggers a tactical response. In the graph for section 9.4, link E→T is a bridge. Removing that link creates two isolated graphs, one of which is a recognizably common satellite studying tactic, T↔F.

Another concept in graph theory, the *cut point*, also may reveal tactics. If a node in a connected graph is deleted, all links into and out from that node are deleted. A cut point is a node in a connected graph that, if deleted, causes the graph to become disconnected. In the graph of section 9.4 (Figure 4), there are two cut points, nodes E and T. Nodes at either end of a bridge often, but not always, are cut points.

Bridges and cut points may not always identify tactics; a "general utility" tactic might be embedded in a graph in a less detectable way. Such tactics might be identifiable by capitalizing on another description of graphs called *strong components*. To reach this concept, we must build on the definition of a *clique*.

A clique is a subgraph within a larger graph in which every node is linked to every other node.[3] That is, a clique is a subgraph in which K_c, one of our constraint statistics, has a value of 0. A strong component relaxes the definition of a clique to allow subgraphs with $K_c \geq 0$, but requires that each node connect directly to every other node or be connected through intermediate nodes within the strong component. Table 3 identifies strong components in the log of Gupta's studying. Studying in Sections 9.1 (see Figure 5) and 9.4 share two strong components, one a three-step tactic involving underlining, making notes, and requesting elaborations and another consisting of the testing-feedback tactic discussed above.

Describing Resemblance Among Actions

Operationally different actions a student performs in STUDY's tutoring environment may or may not trace different forms of cognitive processing. Consider the actions of underlining and note making. These seem distinct. However, each could trace the same cognitive processing, for example, a condition-action rule that discriminates important information followed by maintenance rehearsal enacted as either rereading while underlining or copying text verbatim to a note. If two observed actions can substitute cognitively

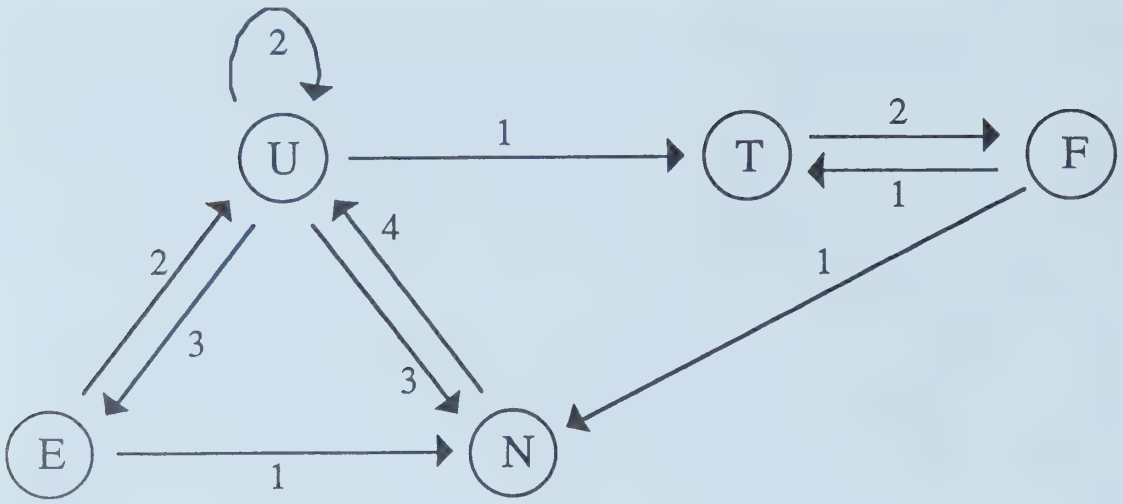


Figure 5. Graph of studying in Section 9.1 {U U N U E U E U N U N U U E N U T F T F N}.

for one another, and if the learner has no preference about alternate expressions for that cognitive processing, the nodes representing each action should have nearly identical relations to other nodes in a graph of the student's studying.

A graph theoretic statistic called structural equivalence can support inferences about whether operationally different traces reflect essentially the same cognitive processing. Structural equivalence compares the roles that two nodes play in a graph by measuring the pair's relationship to every other node in their graph. For each pair of nodes i and j, the distance between them is calculated relative to all other nodes, k, in the graph.

$$d_{ij} = \sqrt{\sum_{\substack{k=1 \\ k \neq i,j}}^{N} (x_{ik} - x_{jk})^2 + (x_{ki} - x_{kj})^2} \ , \qquad (3)$$

where i, j, and k are nodes in a graph, and x_{ij} is the entry in the i^{th} row, j^{th} column of the graph's weighted adjacency matrix. The smaller the structural equivalence coefficient, the more two nodes resemble one another in terms of how those nodes relate to all other nodes. A limiting value of 0 indicates perfect substitutability.

Table 4 shows structural equivalence coefficients for each pair of nodes in Gupta's ALL graph computed using the weighted adjacency matrix. The statistics suggest that actions E (elaborations) and N (notes) function similarly with

Table 3
Strong Components in Gupta's Log

Section	*9.1*	*9.2*	*9.3*	*9.4*	*9.5*
Strong components	[E, U, N] [T, F]	[E, U]	[T, F, N]	[E, U, N] [T, F]	[E, U] [T, F, N]

respect to all the other nodes in the graph, as do T (generating a test question) and F (feedback). We interpret that E and N are both elaborations playing the same general role in Gupta's cognitive engagement, but they differ mainly in their source of information: E is an elaboration that STUDY would provide, whereas N is an elaboration the student generates. Regarding T and F, a similar interpretation is implied. By generating a question about material studied, Gupta creates internal feedback about qualities of her understanding. Feedback that STUDY provides differs in the origin of its information, but it serves essentially the same function as self-generating a question: both actions seek feedback to guide further studying.

A researcher may sometimes be interested in measuring the degree of disresemblance between actions. For instance, suppose a student is being trained to use a study strategy. Plotting resemblance statistics across training sessions describes intraindividual changes—development, functional fixedness, adaptation—in the student's cognitive processing as training progresses. Used in interindividual comparisons, resemblance statistics compared across students can characterize differences in cognitive processing routines.

Describing Resemblance Among Cognitive Tactics and Strategies

As well as comparing single actions that trace features of cognitive processing, research also needs to compare the degree of resemblance or congruence that tactics and strategies have to one another. This would allow comparing interindividual differences in tactics and strategies, as well as intraindividual changes in a student's tactics and strategies when studying different kinds of text or preparing for different kinds of examinations. The graph theoretic statistic multiplicity serves this purpose. For two (or more) graphs that have in common one set of possible nodes, multiplicity measures on a scale from 0 to 1 the number of links the graphs have in common normalized relative to the number of possible links,

$$M = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} x_{ij} \cdot y_{ij}}{N^2}, \quad i \neq j, \tag{4}$$

Table 4
Structural Equivalence of Nodes in the ALL Graph Representing Gupta's Studying

Using the weighted adjacency matrix

	E	*F*	*N*	*T*	*U*
E	.00	16.34	4.69	16.67	7.28
F		.00	13.89	2.24	17.55
N			.00	13.34	10.15
T				.00	17.52
U					.00

Table 5
Multiplicity Values for Sections in Gupta's Log

	9.1	9.2	9.3	9.4	9.5
Section 9.1		.15	.20	.40	.35
Section 9.2			.05	.15	.15
Section 9.3				.25	.30
Section 9.4					.30
Section 9.5					

where x_{ij} is the entry in the i^{th} row j^{th} column of the adjacency matrix for graph X, y_{ij} is the entry in the i^{th} row j^{th} column of the adjacency matrix for graph Y, and N is the number of nodes in each graph.

Again, the question arises, What canonical graph should be used to normalize the statistic? As discussed earlier, we replace N^2 in equation 4 with the number of followed-by links possibly observed in Gupta's data (N=20). Using this adaptation, multiplicity values presented in Table 5 compare the structure of all pairwise sets of graphs tracing Gupta's studying normalized relative to a completely unconstrained canonical ALL graph defined by Gupta's presumptions (i.e., 20 possible links).

The most similar graphs are those for Sections 9.1 and 9.4 (see Figures 4 and 5). Visually these graphs suggest that Gupta used nearly identical studying strategies and tactics. Only the links between the subgraphs [U, E, N] and [T, F] differ across these two sections of the text. Despite differences in the frequency of individual studying acts, there is strong resemblance of Gupta's studying across these sections.

We also want to directly compare pairs of graphs having a common set of possible nodes to one another rather than normalizing them relative to a canonical graph. The question arising in this situation is how to define the normalizing term in equation 4. Our proposal is to sum over the pair of graphs the number of distinct links observed.

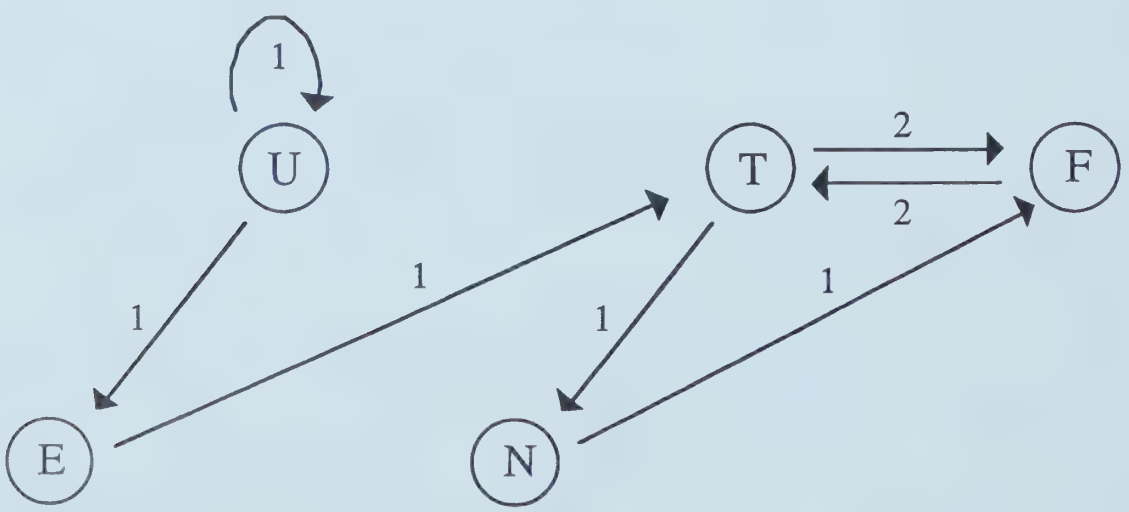


Figure 6. Graph of studying in Section 9.3 {U U E T N F T F T F}.

Table 6
Normalized Multiplicity Values for Gupta's Log

	9.1	9.2	9.3	9.4	9.5
Section 9.1	1.00	$\frac{3}{10}$=.30	$\frac{4}{13}$=.37	$\frac{8}{11}$=.73	$\frac{7}{13}$=.54
Section 9.2		1.00	$\frac{1}{9}$=.11	$\frac{3}{9}$=.33	$\frac{3}{10}$=.30
Section 9.3			1.00	$\frac{5}{11}$=.45	$\frac{6}{11}$=.55
Section 9.4				1.00	$\frac{6}{13}$=.46
Section 9.5					1.00

By the first measure of graph multiplicity in Table 5, Sections 9.3 (see Figure 6) and 9.5 were as comparable as Sections 9.4 and 9.5, M=.30 in both cases. By our alternative measure, Sections 9.3 and 9.5 are more similar to each other, M=.55 versus M=.46 comparing Sections 9.4 and 9.5. This variation is explained by relationships among T, F and N: these actions interact partially in Section 9.3 but not at all in Section 9.4. We intuit that our adapted measure of graph multiplicity is a better gauge of the similarity of graphs and, correspondingly, how Gupta studied in different sections.

Conclusion

We propose that, fundamentally, individual differences exist as cognitive acts that learners perform over the course of engaging with a task. Many, perhaps most, cognitive acts are difficult for a learner to inspect, report, or otherwise reveal because they are: complex, automatic cognitive procedures that can be decomposed only with difficulty and at cost to smooth and effective execution (Anderson, 1983; McKoon & Ratcliff, 1992); tacit (Schommer, 1990; Winne & Butler, 1994); or simply inexpressible with terms in the learner's lexicon. Facing these problems, traditional approaches to studying individual differences such as questionnaires about learning styles and measures of global achievement have aggregated up. Such data obscure individual differences at their source.

A decade ago, Winne (1982) argued that data representing individual differences should be gathered at their source level and examined at that level

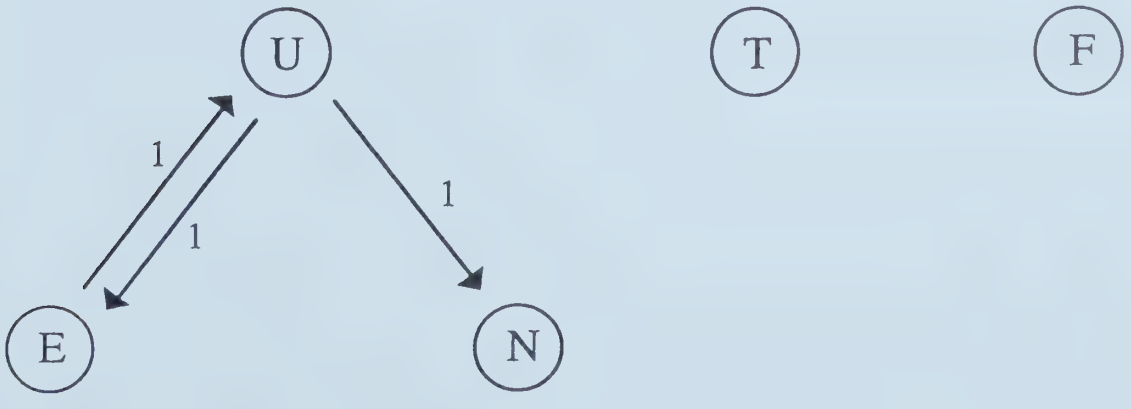


Figure 7. Section 9.2 {U E U N}.

before aggregating up. He argued that analyzing fine-grained traces of cognition would sharpen pictures of how learners express intra- and interindividual differences, and provide more penetrating grounds for examining how individual differences affect the course of learning and its ultimate achievements. The adaptive learning system STUDY was designed in part as a tool for gathering trace data. The question remaining, however, was: How can trace data be analyzed to reveal cognitive acts and patterns that constitute individual differences?

Graph theoretic concepts are one source of techniques for addressing this question. In a task like studying, descriptors and statistics introduced in this article are used to (a) identify units of cognition and patterns of cognitive engagement, and (b) measure the degree to which traces of single cognitive units as well as cognitive tactics resemble one another. Coupled with traces of cognition that are carefully designed to reflect theoretically important cognitive events, these measures can characterize features that give rise to individual differences, namely, conditional knowledge, attention, discriminated action, and goals.

It is important to recognize limitations in this approach to investigating individual differences in tasks such as studying that entail complex cognition. A sequence of traces reflects not only knowledge and competence a learner brings to task, but also the interaction between these and the information studied. For example, the number of study actions and the pattern of studying in Section 9.2 (see Figure 7) suggest that Gupta was quite inactive relative to other sections. Does this indicate lack of comprehension, perhaps occasioned by low prior knowledge, or low levels of strategy use? Section 9.2 was an extended example of applying statistical techniques covered in previous chapters of the text. Although graph theoretic concepts and statistics (see Tables 2 and 3) identified a difference in studying, other information about what was being studied must be considered to grasp how the difference makes sense. This reflects an important principle, that interpreting a statistic is inherently bound up with what raw data reflect. We continue to explore means by which to accommodate the situational qualities of trace data. Although this will complicate matters, we expect that fine-grained studies of such person-environment interactions will also spur newer and more valid theories of individual differences called for by Corno and Snow (1986).

Acknowledgment

This research was supported by a grant to the first author from the Social Sciences and Humanities Research Council of Canada (#410-92-0239).

Notes

1. These data cannot be regarded as providing a sufficient basis for theorizing about actual study strategies.
2. Presumptions about useful or permissible studying activities correspond to decisions a designer would have option to operationalize as rules governing a STUDY tutorial. For example, STUDY can be designed to provide feedback in relations other than a one-to-one correspondence to a specific test question. A designer might supply feedback to a student whenever a particular studying action is carried out, such as alerting the student that text just underlined is a minor detail rather than contributing to the text's main theme or message.
3. The requirement that a node be related to other nodes insures that a node with a self-loop cannot create a clique of size one.

References

Anderson, J.R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Butler, D., Winne P.H., & McGinn, M. (n.d.). *A model of feedback in the self-regulation of learners*. Unpublished manuscript.

Corno, L., & Snow, R.E. (1986). Adapting teaching to individual differences among learners. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 605-629). New York: Macmillan.

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Genest, M., & Turk, D.C. (1981). Think-aloud approaches to cognitive assessment. In T.V. Merluzzi, C.R. Glass, & M. Genest (Eds.), *Cognitive assessment* (pp. 233-269). New York: Guilford Press.

Glover, J.A. (1989). The testing phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*, 392-399.

Goldman, S.R., & Saul, E.U. (1990). Flexibility in text processing: A strategy competition model. *Learning and Individual Differences, 2*, 181-219.

Graham, S. (1991). A review of attribution theory in achievement contexts. *Educational Psychology Review, 3*(1), 5-39.

Guzdial, M. (1993). *Deriving software usage patterns from log files*. Unpublished manuscript.

Haller, E.P., Child, D.A., & Walberg, H.J. (1988). Can comprehension be taught? A quantitative synthesis of "metacognitive" studies. *Educational Researcher, 17*(9), 5-8.

Howell, D.C. (1987). *Statistical methods for psychology* (2nd ed.). Boston: PWS-Kent.

Johnson, L.L. (1988). Effects of underlining textbook sentences on passage and sentence retention. *Reading Research and Instruction, 28*(1), 18-32.

Kierwa, K.A., DuBois, N.F., Christian, D., McShane, A., Meyerhoffer, M., & Roskelley, D. (1991). Note-taking functions and techniques. *Journal of Educational Psychology, 83*, 240-245.

Luchins, A.S., & Luchins, E.H. (1959). *Rigidity of behavior: A variational approach to the effects of einstellung*. Eugene, OR: University of Oregon Books.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*, 440-466.

Peterson, S.E. (1992). The cognitive functions of underlining as a study technique. *Reading Research and Instruction, 31*(2), 49-56.

Schommer, M. (1990). Effects of belief about the nature of knowledge on comprehension. *Journal of Educational Psychology, 82*, 489-504.

Walczyk, J.J., & Hall, V.C. (1989). Effects of examples and embedded questions on the accuracy of comprehension self-assessments. *Journal of Educational Psychology, 81*, 435-437.

Winne, P.H. (1982). Minimizing the black box problem to enhance the validity of theories about instructional effects. *Instructional Science, 11*, 13-28.

Winne, P.H. (1992). State-of-the-art instructional computing systems that afford instruction and bootstrap research. In M. Jones & P.H. Winne (Eds.), *Foundations and frontiers of adaptive learning environments* (pp. 349-380). Berlin: Springer-Verlag.

Winne, P.H. (1993a, May). *Frontiers of research on academic learning*. Paper presented to the Canadian Psychological Association, Montreal.

Winne, P.H. (1993b). A landscape of issues in evaluating adaptive learning systems. *Journal of Artificial Intelligence in Education, 4*, 295-318.

Winne, P.H., & Butler, D.L. (1994). Student cognition in learning from teaching. In T. Husen & T. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed.). Oxford, UK: Pergamon.

Winne, P.H., Jones, M., & Field, D. (1992, June). *STUDY: A tool for authoring adaptive learning environments and for advancing instructional research*. Paper presented to the 2nd International Conference on Intelligent Tutoring Systems, Montreal.

Winne, P.H., Jones, M., Field, D., & Nesbit, J.C. (n.d.). *Instructional research with STUDY*. Unpublished manuscript.

Zimmerman, B.J., & Schunk D.H. (Eds.). (1989) Self-regulated learning and academic achievement. Springer-Verlag: Berlin

The Alberta Journal of Educational Research *Vol. XL, No. 2, June 1994, 195-211*

W. Todd Rogers
University of Alberta

and

David J. Bateson
University of British Columbia

Verification of a Model of Test-taking Behavior of High School Seniors

Thirty-six test-wise and 41 test-naive high school seniors were individually tested using a 14-item test designed to assess their abilities to apply selected test-wiseness principles (Millman, Bishop, & Ebel, 1965) according to a proposed model of test-taking behavior. Students' approaches to their answers were recorded and analyzed. Overall, the results (a) suggest that before students can profitably apply test-wise skills, they must first possess knowledge about the content of the stem and/or options to eliminate an option or options or to take advantage of item cues, and (b) support the proposed model of test-taking behavior.

Trente-six élèves du secondaire ayant développés l'art d'écrire des examens et 41 élèves du même niveau n'ayant pas maîtrisé cet art subirent des épreuves individuels au moyen d'un test à 14 items. Ce test a été conçu pour évaluer leurs habiletés à appliquer certains principes de l'art d'écrire un examen (Millman, Bishop, & Ebel, 1965) d'après un modèle proposé pour déterminer le comportement des élèves durant un examen. La façon que les élèves abordaient les réponses fut notée et analysée. En somme, les résultats suggèrent (a) qu'avant de pouvoir utiliser profitablement les habiletés et le savoir faire pour répondre aux questions d'un test, les élèves doivent d'abord posséder une certaine connaissance au sujet du contenu de la souche de la question et/ou des options pour éliminer une option ou des options ou pour utiliser des indices inhérentes à la question et (b) que les résultats appuient le modèle proposé des comportements en testing.

> STANDARD 3.11—When test-taking strategies that are unrelated to the constructs or content measured have been found to influence test performance significantly, these strategies should be explained to test takers before the test is administered either in an information booklet or, if the explanations can be made briefly, along with the test directions. The use of such strategies by all test takers should be encouraged if their effect facilitates performance and discouraged if their effect interferes with performance. (*Primary*). (American Psychological Association, 1985, pp. 27-28)

Test-wiseness is a cognitive ability or set of skills that a test taker can use to improve a test score no matter what the content area of a test (Benson, 1988;

Todd Rogers is a professor in the Department of Educational Psychology and Director of the Centre for Research in Applied Measurement and Evaluation. He specializes in measurement and evaluation.
David Bateson is an associate professor in the Department of Mathematics and Science Education and a research associate with both the Educational Measurement Research Group and the Centre for Applied Studies in Evaluation. He specializes in science education, measurement, and evaluation.

Sarnacki, 1979). If a test taker possesses test-wiseness, and if the examination contains susceptible items, then the combination of these two factors can result in an improved score; in contrast, a student low in test-wiseness will tend to be penalized every time he or she takes a test that includes test-wise components.

A number of authors have offered definitions of test-wiseness (Diamond & Evans, 1972; Ebel & Damrin, 1960; Gibb, 1964; Stanley, 1971; Thorndike, 1951; Vernon, 1962), but for the purposes of this study the definition proposed by Millman, Bishop, and Ebel (1965) was adopted: "a subject's capacity to utilize characteristics and formats of the test and/or test taking situation to receive a high score" (p. 707). Basically, then, test-wiseness encompasses both the method of measurement (flawed test items that provide test-wise cues) and characteristics of the test taker (a cognitive ability or set of abilities that an examinee might employ in any testing situation regardless of the content measured).

To further explicate the construct of test-wiseness, Millman et al. (1965) included with their definition a taxonomy of test-wiseness principles that has served as the general framework for further studies of test-wiseness. Briefly, this taxonomy is organized into two categories. Part I contains elements applicable in most testing situations and that are independent of the test maker or test purpose. If employed, these strategies will help examinees avoid losing points for reasons other than lack of knowledge of the content tested. The principles listed in Part II of the taxonomy may prove beneficial when the test taker has knowledge of particular test making behaviors or knowledge of particular testing practices gained from past experiences with tests similar in purpose and format. For more detail, the reader is referred to Millman et al., (1965) and Sarnacki (1979).

Model of Test-wise Test Taking Behavior

Displayed in Figure 1 is a proposed model of the test taking behavior of skilled test takers. Based on the work of Brown (1980; 1987), Flavell (1979), and Schuell (1986), and an earlier model proposed by Smith (1980), the model reflects various routes a skilled examinee may take to determine what option to select on a multiple-choice item, and an "executive" governor for selecting which routes are used. Briefly, the model suggests that the cognitions include:

1. a cognitive monitor that controls which abilities and skills are going to be engaged to answer the item under consideration;
2. knowledge, abilities, and skills relevant to the content or trait being measured;
3. knowledge of test-wiseness principles; and
4. the response (selection and record of choice).

The proposed process follows a defined path. First, the test taker reads the stem of a multiple-choice item and then attempts to recognize, using knowledge about the perceived content being tested, what he or she believes to be the correct answer from among the options listed. If the answer is not found, an unskilled test taker will either simply guess from among the options presented or omit the question entirely. In contrast, a skilled test taker, by way of his or her cognitive monitor for testing and partial knowledge about the content being measured (including that contained in the item's options), will next apply the set of test-wiseness principles he or she possesses, working cyclically through the elements of the set for a test-wiseness element-item cue match.

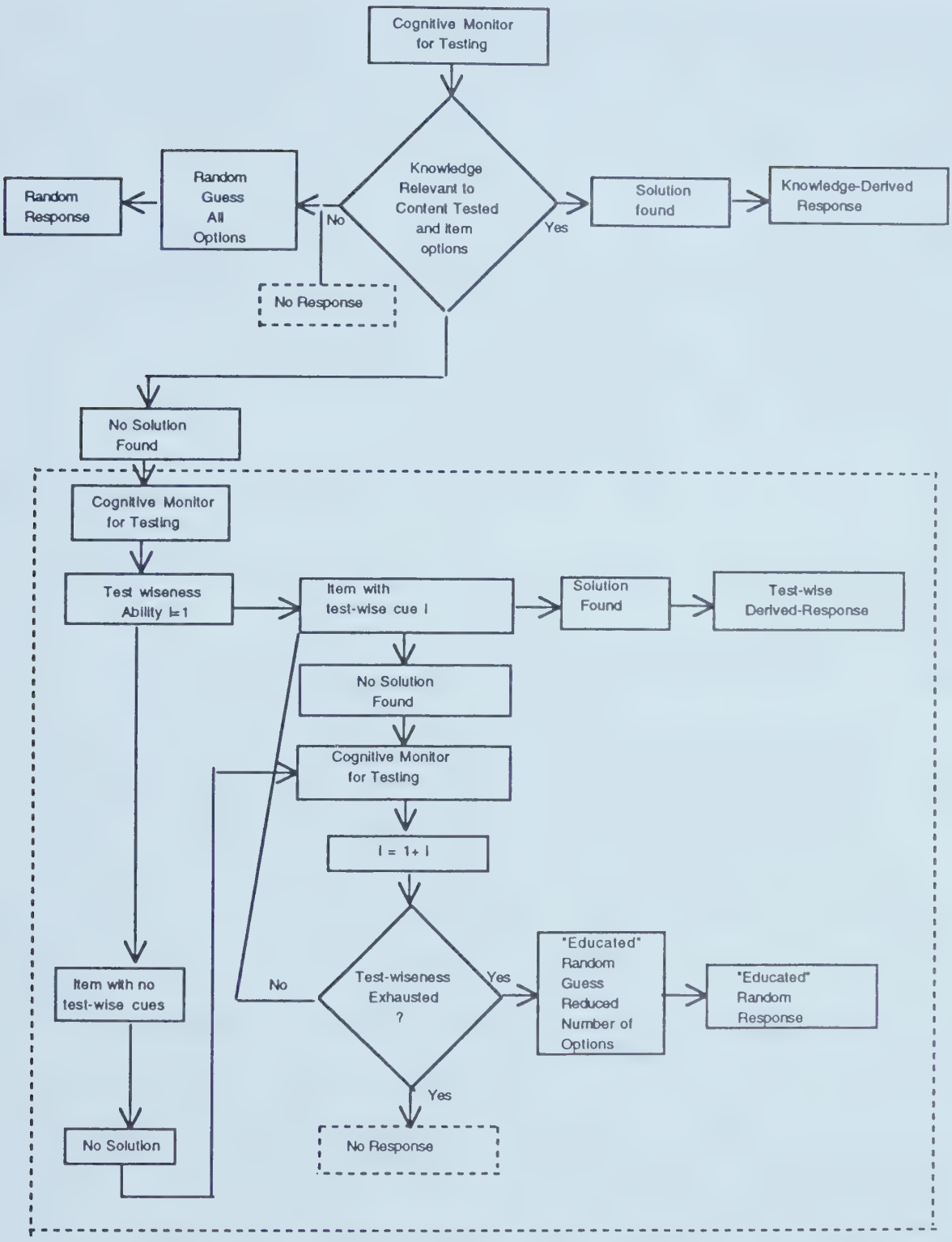


Figure 1. Proposed model of test-taking behavior of skilled test-takers.

When a match is made, the cycle is terminated and a test-wise (as opposed to a pure knowledge) response is recorded. In the case of no match, either because there are no item test-wise cues or because the test-taker has exhausted all of his or her test-wiseness strategies, the skilled test taker will probably make an "educated" random response. Clearly then, according to this model, other characteristics of the skilled test taker in addition to knowledge of the content being tested enter into a high test-wise student's final test score.

The study described and reported in this article was designed to assess the validity of this test-taking model. Completed as part of a larger study of the

impact of test-wiseness on the performance of high school students in British Columbia on school leaving examinations (Rogers & Bateson, 1991), the intent was to gain a more complete understanding of test-wiseness and the way and extent to which high school seniors actually employ test-wiseness when responding to multiple-choice items.

Method

Test of Test-wiseness

The Test of Test-wiseness (TTW) developed and used in this study was divided into two sections. Section 1 contained 24 four-option, multiple-choice items evenly distributed across four of the test-wiseness elements listed in the taxonomy outlined by Millman et al. (1965). The four elements included:

1. Three deductive reasoning strategies:

   ID1—eliminate options known to be incorrect;

   ID2—choose neither or both of two options which imply the correctness of each other;

   ID3—choose neither or one of two-options, one of which, if correct, would imply the incorrectness of the other; and

2. One cue-using strategy:

   IIB4—recognize and use similarities between the stem and the options.

The selection of these four elements was based on the most frequently occurring item faults identified in the provincial examinations administered during the previous year in British Columbia. The 24 items used were selected from test-wiseness tests developed earlier by Gibb (1964), Millman (1966), and Slakter, Koehler, and Hampton (1970). To provide face validity, the content of the items, unfamiliar to the students, was distributed approximately evenly across the content areas of four grade 12 courses with the largest student enrollments—English, algebra, history, and biology (see Table 1). However, the actual content tested by each item was deliberately unfamiliar to the students so as to rule out the possibility that correct answers would be attributable to student knowledge alone.

The intent of the first section was to measure the degree to which students were able to use deductive reasoning to answer questions about content unfamiliar to them, but that were flawed in some way. Group administered, the instructions to the students were:

> This is a test of test-wiseness which measures some of the abilities needed to do well on tests. Many of the questions are about things you may not have studied. However, there are test-taking strategies which can be used to figure out what to do when faced with such questions.
>
> *For example:*
> *The greatest advantage of using slent in the manufacture of steel is that slent makes steel*
>
> a. transparent.
> b. stainless.
> c. heavy.
> d. rubbery.

Table 1
Description of Test of Test-wiseness

Test-wiseness Principle	*Content Area*				
	English	*Math.*	*History*	*Biology*	*Total*
Section 1					
Absurd options(1D1)	2	1	2 (2)	1	6(2)
Similar options(1D2)	2	1	1 (1)	2 (1)	6(2)
Opposite options(1D3)	2	2	1	1 (2)	6(2)
Stem-option (IIB4)	1	1 (1)	2 (1)	2	6(2)
Total	7	5 (1)	6 (4)	6 (3)	24(8)
Section 2					
True-False	1	- (1)	-	1 (1)	2(2)
Multiple-Choice					
Guessing	1	1 (1)	1 (1)	1	4(2)
Non-guessing	1	1	1 (1)	1 (1)	4(2)
Total	3	2 (2)	2 (2)	3 (2)	10(6)

Note: Numbers in parentheses refer to the number of items on the interview form.

> Using test-wiseness strategies, options 'a' and 'd' can be eliminated because they are clearly not correct (steel is not transparent, nor is it rubbery). Therefore, either 'b' or 'c' is the correct answer. Now we stand a better chance of guessing the correct answer for we have narrowed the number of possible options down to two from four.
>
> *Please be sure to follow the specific instructions for each of the following sections.*
>
> *Section 1: Suggested Time: 20 minutes*
>
> *INSTRUCTIONS: For each question, select the BEST answer and record your choice on the answer sheet provided. Each question is worth one mark. There will be no correction for guessing.*

Section 2 involved the use of guessing whenever the chance of profiting was positive. Following Millman (1966), the penalty system for this section was such that test-wise behavior should lead examinees to guess on true-false questions and those multiple-choice questions for which more than one option could be eliminated:

> INSTRUCTIONS: For each question in this part of the test, select the BEST answer and record your choice on the answer sheet provided.
>
> Each question answered correctly is worth 5 points. Zero points will be given for each omitted question. Two points will be *deducted* for each question answered incorrectly.

Of the eight multiple-choice questions asked, four were such that guessing was profitable; the options on the remaining four questions were such that it was

not possible to clearly eliminate an option (see Table 1). The following two examples illustrate these two types of questions:

Example 1. Guessing is profitable

30. The human skin
    a. possesses a Malpighian layer.
    b. is always brown.
    c. cannot regenerate.
    d. is uncolored.
    e. is composed of cuboidal epithelium.

Example 2: Guessing is not profitable

32. What is the value of 1n sin 8º15′ to five decimal places?
    a. .14321
    b. .14332
    c. .14349
    d. .14352
    e. .14361.

Scoring. The system used to score the TTW was designed to reflect the use of the test-wiseness elements assessed. Modelled after Millman (1966), the items were scored as follows:

Section 1. One point for

ID1: correct answer;
ID2: either of the two nonsimilar options;
ID3: either of the two opposite options; and
IIB4: the cued option.

Thus the range of possible scores was zero to 24; a high score reflected frequent use of test-wiseness reasoning.

Section 2. Two points were awarded for every true-false and "guessing" multiple-choice item answered, regardless of whether the answer was correct. Three points were awarded for each of the "nonguessing" multiple-choice items not answered. Thus the maximum score was 24; a high score reflected appropriate use of a guessing strategy.

Interview form. A second, shorter form of the TTW was developed and administered on an individual basis to identify the actual processes students follow when answering questions with unfamiliar content. This form consisted of eight four-option, multiple-choice items evenly distributed across the four test-wiseness elements included in the longer, group administered form and two true-false and four five-option multiple-choice items to measure guessing strategy (see Table 1). Questions were obtained from Millman (1966) and the provincial biology examination administered the previous year.

Subjects

The subjects were students who wrote the provincial examination set for English 12 in June 1989. All grade 12 students must write either English 12 or Communications 12 in order to graduate. English 12, intended for students who are planning or wish to leave open the opportunity to pursue some form of tertiary education, is more popular, enrolling between five and six times as many students as Communications 12. Further, students who take English 12

also elect additional courses with provincial examinations so as to meet postsecondary entrance requirements. Thus, by selecting students in English 12, samples of adequate size would be realized for each of the five remaining provincial tests considered in the larger study.

The students attended 10 public schools selected to provide provincial representation. Three of the schools—two large and one medium sized—were located in the southwest region of British Columbia. The others were distributed throughout the province: one large, one medium, and one rural school in the south central region; one medium and one rural school in the north central region; and two rural schools on the north coast. The students attending these schools were from diverse socioeconomic backgrounds, reflected the gender distribution and major ethnic groups in the province, and represented the full range of achievement. Summarized in Table 2 are the gender and ethnic distributions for the English 12 student sample (n=736) and the corresponding population distributions. Comparison of the sample values with the corresponding population parameters confirms the representativeness of the student sample (see Table 2).

Table 2
Gender, Age, and Ethnicity

Variable	*English 12 Sample (N=736)* *f*	*p*	*Population Percentage*
Gender			
Male	343	46.4	47.9[a]
Female	393	53.2	52.1
Age (Year of Birth)			
1967	1	0.1	n.a
1968	4	.5	n.a
1969	5	.7	n.a
1970	87	11.8	n.a
1971	618	84.0	n.a
1972	14	1.9	n.a
Ethnicity			
English	430	58.2	61.8[b]
French	21	2.8	2.4
Indigenous	11	1.5	2.1
East Indian	30	4.1	n.a.
Chinese	36	4.9	4.0
German	44	6.0	5.2
Italian	24	3.2	1.6
Japanese	8	1.1	n.a.
Other	132	17.9	n.a.

Note: Percentages do not total 100% due to omitted responses. n.a. = not available.
[a]Percentage of total Grade 12 provincial population.
[b]Percentage of total provincial population in 1986 (Sullivan, 1988, p.27).

Data Collection

The student data considered in this study were collected in two stages. At stage 1, an initial sample of 936 students in the 10 schools completed the TTW during an English 12 class at the beginning of May. The TTW was administered by the authors assisted by the teachers when more than one English 12 class was scheduled at the same time. The testing time required was 25 minutes.

Approximately two weeks later a subsample of the highest scoring (test-wise; n=36) and lowest scoring (test-naive; n=41) students on the nonguessing subtest of the TTW were individually tested to ascertain the response strategies the students used when answering each item in the interview form of the TTW. In order to reduce costs, this subsample was restricted to students enrolled in the six schools located in the southwest and south central regions of the province. The students were asked by the interviewer, a member of the research team, to "think aloud," to "describe and explain what you are doing," or "how did you get your answer?" for each of the 14 questions. Each interview lasted 15 to 20 minutes. Generally, the test-wise students required less time than the test-naive students.

Results

Description of Interview Samples

As previously described, the subjects selected for this part of the study formed two subsamples: 36 test-wise students whose scores on Section 1 of the TTW were greater than 17, and 41 test-naive students whose scores on Section 1 of the TTW were less than 11. The cut-off scores are approximately one standard deviation above and below the mean on Section 1 of the TTW for the initial sample of 936 students ($\bar{X}$ = 13.9; s=2.80). The percentages of males and females in both groups were approximately equal: 38.9% of the test-wise sample and 41.5% of the test-naive sample were male, and 61.1% of the test-wise sample and 58.5% of the test-naive samples were female. Nearly equal percentages of students in both groups, 52.8% and 53.7%, reported that they were of British/European ancestry. The next most frequently reported ethnic background was "Other," 36.1% and 19.5%. The difference between these two percentages is mainly accounted for by smaller percentages of Aboriginal (0.0% vs 4.9%) and East Indian (5.6% vs 12.2%) students in the test-wise sample than in the test-naive sample.

Solution Strategies Used

Turning now to the strategies used by the students in these two samples, the percentages of students in each sample who correctly responded to each item included in Section 1 of the Interview Form are reported in Table 3.

The results are organized in terms of the strategies that, if used properly, would lead to an increased probability of a correct response. Table 3 reveals that the differences between item percentages in favor of the test-wise sample are significant on two items (4 : z = 2.65; 8 : z = 2.35; $p < .05$). Despite the lack of statistical significance on other items in Section 1 with large differences (items 1, 2, 5, 6, 7), there are discernible differences between the solution strategies the test-wise students reported they used and the solution strategies the test-naive students reported they used. These are described next in terms of test-wise principles most frequently used by the students when answering each item.

Table 3
Performance of Interview Samples on Interview Form, TTW
Section 1

Test-wiseness element	*Item #*	*Sample*			
		Test-wise (n = 36)		*Test-naive (n = 41)*	
		n	*%*	*n*	*%*
ID1	1	19	52.8	14	34.1
	8	15	41.6	6	14.6*
ID2	4	27	75.0	20	48.8*
ID3	2	16	44.4	14	34.1
	5	25	69.4	22	53.6
IIB4	3	19	52.8	21	51.2
	6	28	77.8	25	61.0
ID1, ID2	7	27	75.0	29	70.7

*$p < .05$

ID1: Option known to be incorrect

1. The compromise between Democrats and Republicans after the post-civil war election of 1876 resulted in
   A. federal aid to Southern railroads.
   B. the entrance of Texas into the Union as a slave state.
   C. a treaty with Joseph Stalin.
   D. the Fugitive Slave Law. (Millman, 1966)

8. The emperor of the ancient Hsin Dynasty who resigned to undertake radical reform was
   A. Saigon.
   B. Wang Mang.
   C. Mao Tse Tung.
   D. Alexander I. (Millman, 1966)

Items 1 and 8 contain incorrect options that, if known by a student to be incorrect, should be eliminated or ignored. Ten (27.8%) students in the test-wise sample identified as wrong and then eliminated the three incorrect options in item 1. An additional four students first said they knew option C was wrong and then pointed out that, because B and D were similar, they too were wrong; a fifth test-wise student first indicated B and D were similar and then correctly guessed A. In contrast, five (12.2%) students in the test-naive sample correctly identified and eliminated the three incorrect options; no student in this sample reported B and D to be similar. Thirteen (36.1%) students in the test-wise sample and nine (22.0%) students in the test-naive sample eliminated one or two options (including, in some cases, A) as incorrect and then guessed among the remaining three or two options. Two (5.6%) students in the test-wise sample reported that they knew the answer (both incorrect), and seven (19.4%) indicated that they simply guessed (two correctly, which agrees with what would be expected by simple chance). The incidence of knowledge and guess-

ing in the test-naive sample was greater: 12 (29.3%) students said they knew the answer (only three were correct), and 15 (36.6%) reported they guessed (three correctly which again agrees with what would be expected by chance alone).

The pattern of solution strategies used by the students to answer item 8 is somewhat similar to that observed for item 1. Although 14 (38.9%) students in the test-wise sample correctly identified and eliminated the three incorrect options in item 8, only three (7.3%) students in the test-naive sample did likewise. Twelve (33.3%) test-wise students and 16 (39.0%) test-naive students eliminated one or two options (including in some cases the correct option) and then guessed among the remaining three or two options. Three additional students, one test-wise and two test-naive, used strategy ID1 to eliminate three options, one of which was the correct option. Two test-wise students first used ID1 to eliminate option D and then selected A by guessing (one student) or citing knowledge ("I know this from a class I took"). Four test-naive students followed the same strategy to eliminate A and/or B and then selected C (3) or D (1) citing knowledge. Two students in the test-wise sample and three in the test-naive sample used the stem-option similarity to select C. Three students in each sample selected their answer (all of which were incorrect) based on knowledge they claimed they had. Last, whereas only one test-wise student said he guessed, eight (19.5%) students in the test-naive sample said they guessed (two correctly).

ID2: Similar Options

4. The sigma effect of the Fahraeus-Lindgvist phenomenon is related to
   A. the flow of liquid through the kidney.
   B. the muscular contractions of the kidney.
   C. the kidney's transmission of fluids.
   D. the diameter of the red cells flowing through the kidney's blood vessels. (Millman, 1966)

As was the case for the two previous items, a greater percentage of students in the test-wise sample than the percentage of students in the test-naive sample correctly took advantage of the existence of similar options in item 4. Twenty-one (58.3%) test-wise students identified A and C as similar and selected their answers from B and D: 16 guessed, two used ID1 to eliminate D, another used ID1 to eliminate B, and two selected D based on knowledge "learned in class." In contrast, five (12.2%) test-naive students identified A and C as similar: four reported they guessed between B and D and one selected D based on previously learned knowledge. The majority of test-naive students, 25 (61.9%), reported they guessed (seven correctly); only nine (25.0%) test-wise students reported they guessed (two correctly). And whereas eight (19.5%) test-naive students said they knew the answer, only one test-wise student said she knew the answer.

ID3: Opposite Options

2. If ribulose biphosphate were removed from a chloroplast, which of the following statements would BEST describe the immediate result?
   A. CO_2 could not enter the Calvin cycle.
   B. ATP could not be produced in the thylakoid.
   C. O_2 could not enter the Calvin Cycle.
   D. Light energy could not be trapped in the grana. (Ministry of Education, 1986)
5. Which of the following would cause an oat seedling to bend to the right?
   A. Auxins placed on the left side of the shoot tip.
   B. Giberellins placed on the left side of the shoot tip.
   C. Cytokinins placed on the right side of the shoot tip.
   D. Auxins placed on the right side of the shoot tip. (Ministry of Education, 1986)

Unlike the previous three items, items 2 and 5 were selected from the previous year's provincial examination in biology. Consequently, although both items contained opposite options, several students, particularly in the test-wise sample, cited the use of previously learned material in arriving at their answer. In contrast, the incidence of guessing was greater in the test-naive sample. Further, a greater number of students used more multiple strategies on items 2 and 5 than they did on the three previous items. On item 2, for example, 20 (55.6%) test-wise students identified A and C as opposites, although only six (14.6%) test-naive students recognized that A and C were opposites. Among the 20 test-wise students, nine first indicated A and C were both incorrect and then selected B or D using knowledge (5) or guessing (4); three first eliminated option D as incorrect, recognized A and C were opposites and then selected A or C using knowledge (2) or guessing (1); and eight recognized A and C as opposites and selected between these options using knowledge (3) or guessing (5). Two of the six test-naive students who said A and C were opposites indicated both were wrong and, citing knowledge learned in class, selected B; one first eliminated D and then guessed between A and C; and three guessed between A and C. Although four (11.6%) test-wise and six (14.6%) test-naive students indicated they used knowledge alone to get their answers, only five (13.9%) test-wise students compared to 22 (53.9%) test-naive students simply guessed their answers.

As might be expected, the pattern of solution strategies used by the students to answer question 5 is similar to that for item 2. Again the majority of test-wise students, 26 (72.2%), recognized that A and D were opposites whereas a minority of test-naive students, 13 (31.7%) did the same. Six of the 26 test-wise students indicated that A and D were wrong: one then eliminated B as incorrect, another used the similarity between the stem and option C (the word "right"), three used knowledge learned in class to pick between B and C, and one, after omitting A and D, guessed between B and C. The remaining 20, all of whom selected either A or D, did so in the following ways: one first eliminated B, recognized A and C were opposites, and then guessed; two used the presence of "right" in the stem, one to select D, the other to select A; and 17 chose between A and C on the basis of knowledge (6) or guessing (11). Of the

13 students in the test-naive sample who recognized A and D were opposites, six indicated these options were incorrect and then selected option C (two used knowledge; one eliminated A as incorrect, one used stem-option similarity, and two guessed); one eliminated C as incorrect, recognized A and D as opposites, and then eliminated D as incorrect; and seven first recognized A and D as opposites and then selected A or D using knowledge (2), stem-option (1), or guessing (3). Again, whereas four (11.1%) test-wise students and six (14.6%) test-naive students indicated they used knowledge alone, only four (11.1%) test-wise students compared to 20 (51.2%) test-naive students simply guessed.

II.B.4: Stem-Option Similarity

3. In 1991 the Tories in control of the House of Commons showed their political influence and democratic spirit when they
   A. enacted the Asquith price-control bill.
   B. vetoed a bill to extend the suffrage privilege.
   C. forced the House of Lords to accept a reduction in its power.
   D. were defeated in an attempt to enact the National Insurance Act. (Millman, 1966)

6. Why is Cavalieri's Principle important in Solid Geometry?
   A. It shows that the surface area of a square of side s is s^2.
   B. It provides contradictions to the principles of Euclid and Gauss.
   C. It is used to prove that two polygons are congruent.
   D. It provides the basis for finding the volume formulae for many solids. (Millman, 1966)

As with the three previous test-wise elements, students in the test-wise sample made greater use of the stem-option cue built into items 3 and 6 than did students in the test-naive sample. Sixteen (44.4%) test-wise students identified a similarity between the stem and one of the options in item 3; 13 (31.7%) test-naive students did the same. Eleven of the 16 test-wise students correctly used the cue to select option C: seven directly, three following elimination of option D as incorrect, and one supported with knowledge learned in class. The remaining five students selected option B, one directly, one following elimination of D, and three supported by knowledge. Of the 13 test-naive students who used the stem-option cue, 10 used it correctly: nine directly and one following elimination of option D. The remaining three students all selected B, one directly and two with knowledge. Nine (25.0%) additional test-wise students used ID1 in determining their answer in much the same way as that observed for items 1 and 8: four students identified A, B, and D as wrong while two others identified B, C, and D as wrong; two identified A and D as wrong and then guessed between B and C; and one eliminated D and then guessed. Four (9.8%) test-naive students followed the same procedures: one eliminated A, B, and D to obtain the correct answer; one eliminated B and D and then guessed; and two eliminated D and then guessed. Eight (22.2%) test-wise students reported they simply guessed their answers (three correctly) in contrast to 18 (43.9%) test-naive students who reported they guessed their answers (eight correctly).

For item 6, 21 (58.3%) students in the test-wise sample used the stem-option similarity cue to obtain their answers, whereas 15 (36.6%) test-wise students

did likewise. Thirteen test-wise students used this cue directly to select D, whereas six first eliminated B and/or C as incorrect before making the association between "solid" and "volume." The remaining two students selected A and C, the latter after eliminating B as incorrect. All 15 test-naive students who used the stem-option cue selected option D, 13 directly and two following elimination of option B. Five (13.9%) test-wise students eliminated the first three options, as did one test-naive student. Two (5.6%) test-wise students indicated they knew the answer (one correct, the other wrong) and four (11.1%) reported they simply guessed. In contrast, seven (17.1%) students in the test-naive sample stated they knew the answer from a previous course (one correct, six incorrect) and 17 (41.5%) said they simply guessed (seven correctly). It is noteworthy that in contrast to the previous items discussed, the number of students in the test-naive sample who reported they guessed their answers to items 3 and 6 is greater than (and not equal to) that expected by chance alone. These results suggest that for these students something else other than chance was at work. When probed, many of these students were not able to articulate how they obtained their answer and concluded "I guess I guessed." It was though they had some sense of the similarity between the stem and option C (item 3) and D (item 6).

7. A major characteristic of natural resources is that they are
   A. sources of energy.
   B. nonrecyclable.
   C. nonrenewable.
   D. unevenly distributed. (Ministry of Education, 1986)

During the test construction phase, it was thought that students high in test-wiseness would recognize the similar options in item 7 and then select from among the remaining two-options and that students lower in test-wiseness would fail to see the similar options and therefore perform less well. However, the solution strategies adopted by students in both samples suggested that the material tested by item 7 was more familiar to the students than was the material tested in the previously discussed items. This seems likely given the frequent discussions of natural resources and their depletion in daily and weekly newspapers and on radio and television. In comparison to the previous seven items, the number of students in the test-naive sample who said they simply guessed their answer, two (4.9%), is significantly smaller, suggesting that they too possessed knowledge to answer this question. Indeed 27 (75.0%) students in the test-wise sample and 29 (70.7%) students in the test-naive sample avoided the two similar options. Eight test-wise students identified B and C as similar and used knowledge (4), deductive strategy ID1 (1), or guessing (3) to select their final answer. Only two test-naive students used similar options, one in conjunction with knowledge and the other in conjunction with ID1. Eight test-wise students employed ID1 to eliminate three options as incorrect. Seven test-naive students did likewise. Two test-wise and two test-naive students eliminated option (C). The two test-wise students then selected A (1) and D (1) using knowledge whereas the two test-naive students both guessed A. One test-wise student used stem-option similarity, citing the similarity between "resources" in the stem and "sources" in A, and another used "specific determiners" to eliminate B and C. Thirteen (36.1%) test-wise

and 28 (68.3%) test-naive students said they knew the answer. Finally, as indicated before, two (4.9%) test-naive students reported they guessed, as did three (8.3%) test-wise students.

Guessing

Following Millman (1966), the items designed to test guessing were scored as follows:

1. if a student selects an answer to an item where guessing is favored, score two. Otherwise score zero.
2. If a student does not select an answer to an item where guessing is not favored, score three. Otherwise score zero.

Given that six of the 10 items were guessing items and four were not, the total possible score was 24. The mean of the test-wise sample on the guessing subtest of the Interview Form was significantly greater than the mean of the test-naive sample (18.0 vs. 14.1; $p < .05$). Clearly, as expected from the performance of the students on the first section, the students in the test-wise sample were better able to identify and eliminate incorrect options on the four multiple-choice questions designed to encourage guessing.

Summary and Conclusions

Listed in Table 4 is a summary of the most frequently used strategies employed by the students in the test-wise sample and the students in the test-naive sample when they responded to the eight items included in Section 1 of the Interview Form, and the percentages of students in both samples who simply guessed their answers. Considered together, the greater use of test-wiseness strategies to answer an item by the test-wise sample and the marked reduction in guessing by the test-naive sample on item 7 with its familiar content illustrate the important role knowledge assumes in the application of the test-wiseness elements considered. Before students can apply a test-wiseness strategy to answer a multiple-choice item for which they do not know the answer and which, due to its flawed character, provides a test-wise cue, they must have some knowledge of the content being tested and that contained in the item's options. That the two samples compared differ in knowledge is supported by the observations of the principals of the schools in which students were interviewed. The principals pointed out that the students classified as test-wise were more academically talented, whereas the students classified a less test-wise or test-naive were among the less capable academic students. They further indicated that students in the test-wise sample were enrolled in a greater number of academic courses and were more likely to enter university than the students included in the test-naive sample.

The solution strategies used by the students also support the model of test-taking behavior presented in Figure 1. Clearly, if students did not have knowledge or believed they did not have knowledge relevant to the content tested, then, as illustrated by the following student responses, they were unable to engage a test-wise strategy in an attempt to answer the item:

Student A, Item 2: Oh my gosh! (laughter) I guessed.

Student B, Item 5: Guess this one too. I know nothing about this one too.

Table 4
Percentage of Use of Test-wiseness Elements

		Sample									
Test-wiseness element	*Item #*	*Test-wise (n=36)*					*Test-naive (n=41)*				
		ID1	*ID2*	*ID3*	*I IB4*	*G*[a]	*ID1*	*ID2*	*ID3*	*IIB4*	*G*
ID1	1	63.9	13.9	-	-	19.4	34.1	-	-	-	36.6
	8	80.6	-	-	5.6	2.7	61.0	-	-	7.3	19.5
ID2	4	8.3	58.3	-	-	25.0	-	12.2	-	-	61.0
ID3	2	8.3	-	55.6	-	13.9	2.4	-	-	14.6	53.6
	5	5.6	-	72.2	8.3	11.1	4.9	-	31.7	4.9	48.8
IIB4	3	36.1	-	-	44.4	22.2	12.2	-	-	31.7	43.9
	6	33.3	-	-	58.3	11.1	7.3	-	-	36.6	41.5
ID1, ID2	7	27.8	22.2	-	2.8	8.3	22.0	4.9	-	-	4.9

Note: Test-wise element counted whether final answer correct or incorrect and whether or not it was used alone or in combination with another.
[a]Simply guessed.

Student C, Item 4: Probably A ... I'm not sure, just guessing. I'm trying to eliminate answers out but I don't know ...

The strategy employed by students who possessed or believed they possessed relevant knowledge, but who did not find an answer using this knowledge alone, clearly followed the sequence shown in the dotted box of Figure 1. Typical of the responses of these students were:

Student D, Item 1: Not C because Stalin has to do with Russia ... A, I guess.

Student E, Item 1: Not C ... Stalin after the civil war. Um ... B and D both had to do with slaves and can't be right. So the answer is A.

Student E, Item 2: D is wrong! Oh, A and C are opposites so one is probably right.

(Probe: Why did you select C?)

I guessed between A and C.

Student F, Item 4: Cross out A and C ... they are the same.

(Probe: Why did you select D?)

It seems most complicated ... that's why.

Student D, after eliminating one option because it was known to him to be incorrect (element ID1), was unable to rule out other options. He then made a random guess. Student E first eliminated option C on item 1 using ID1. She then treated B and D as similar and pointed out they could not be correct (ID2). She then had a match and selected the test-wise response A. On Item 2, she again eliminated one option, D, as wrong (ID1). She then recognized that A and C were opposites and decided one was correct. Not having sufficient knowledge to decide, she guessed (ID3). Student F avoided options A and C on item

4 because they were similar (ID2). She then selected D on the basis of its complexity. The knowledge of these three students, although not sufficient to simply answer an item, coupled with their test-wiseness abilities, allowed them to produce a response after eliminating at least one option as incorrect.

The difference in approach observed between the test-wise and test-naive students can be explained by differences in cognitive monitoring. Although all six students quoted above initially experienced "cognitive failure" (Garner, 1990, p. 518), the test-wise students also experienced "metacognitive success" (p. 518). Unlike test-naive students who more frequently guessed their answers from among the options of an item they did not know, test-wise students were better able to take deliberate follow-up actions, integrating the partial content or substantive knowledge they possessed with their ability to take advantage of test-wiseness item cues or clues to systematically eliminate one or more of the options (Brown, 1987). Consequently the total scores of such students on a test containing flawed items will be spuriously inflated:

$$X_{tot} = X_{kd} + X_g + X_{twd/pk} + X_{er/pk} > X_{kd},$$

where

X_{tot} = the total number of correct responses.

X_{kd} = the number of correct knowledge derived responses, and the score a test was initially designed to illicit,

X_g = the number of correct random responses,

$X_{twd/pk}$ = the number of test-wise (integration of partial knowledge and test-wiseness strategies) derived responses, and

$X_{er/pk}$ = the number of "educated" random responses following application of test-wiseness.

Thus it appears that the effective application of test-wiseness reasoning strategies is dependent on some partial knowledge. This partial knowledge, although inadequate to respond to a test item solely on the basis of this knowledge, is sufficient when coupled with knowledge of the test-wiseness principles to increase the probability of correctly responding to items susceptible to test-wiseness. Students with low content knowledge but test-wise knowledge and students with partial knowledge but low test-wise knowledge will perform less well than students who possess both on such items (compare Garner, 1990, p. 520).

These results suggest that a test taker who possesses test-wiseness—partial knowledge about what an item measures and ability to take advantage of item cues—will have a greater probability of correctly answering a test-wise susceptible item than will a student low in test-wiseness. Thus a potential validity problem may exist when one attempts to interpret the meaning of a test score. If the partial knowledge component of test-wiseness is considered relevant when enhanced by a subject's capacity to utilize clues (flaws) present in items, then the interpretation of the total score as a valid indicator of achievement may be justified. But, to the extent this partial knowledge is considered not relevant, the interpretation of the test score will be confounded by construct-irrelevant easiness (Messick, 1989, p. 34). Further, if the intent is to compare

students based on their total scores, the difference between high and low scores on an examination containing test-wiseness clues will be larger than if only knowledge, or the lack thereof, was contributing to this difference. If test scores can be influenced by test-wiseness, then individuals involved in test development, administration, and/or interpretation need to carefully consider the construct of test-wiseness and how it affects scores.

Acknowledgments

The research was supported by the Social Sciences and Humanities Research Council, Canada. Reprinted with permission from the *Journal of Experimental Education, 59*(4), 331-350.

References

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Benson, J. (1988). The psychometric and cognitive aspects of test-wiseness: A review of the literature. In M.H. Kean (Ed.), *Test-wiseness* (pp. 1-25), Bloomington, IN: Phi Delta Kappa.

Brown, A.L. (1980). Metacognitive development and reading. In R.J. Spiro, B.C. Bruce, & W.F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 453-481). Hillsdale, NJ: Erlbaum.

Brown, A.L. (1987). Metacognition, executive control, self-regulation and other even more mysterious mechanisms. In F.E. Wienert & R.H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65-116). Hillsdale, NJ: Erlbaum.

Diamond, J.J., & Evans, W.J. (1972). An investigation of the cognitive correlates of test-wiseness. *Journal of Educational Measurement, 9*, 145-150.

Ebel, R.L., & Damrin, D.E. (1960). Tests and examinations. In C.W. Harris (Ed.), *Encyclopedia of Educational Research* (3rd ed.). New York: Macmillan.

Flavell, J.H. (1979, October). Metacognition and cognitive monitoring. *American Psychologist, 34*, 906-911.

Garner, R. (1990). When children and adults do not use learning strategies: Toward a theory of settings. *Review of Educational Research, 60*(4), 517-529.

Gibb, B.G. (1964). *Test-wiseness as a secondary cue response.* Unpublished doctoral dissertation, Stanford University. (University Microfilms, No. 64-7643)

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education and Macmillan.

Millman, J. (1966). *Test-wiseness in taking objective achievement and aptitude examinations.* Final Report, College Entrance Examination Board.

Millman, J., Bishop, H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*, 707-726.

Ministry of Education. (1986). *Biology 12.* Victoria, BC: Author.

Rogers, W.T., & Bateson, D.J. (1991). The influence of test-wiseness upon performance of high school seniors on school leaving examinations. *Applied Measurement in Education, 4*, 159-183.

Sarnacki, R.E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research, 49*, 252-279.

Schuell, T.J. (1986). Cognitive conceptions of learning. *Review of Educational Research, 56*, 411-436.

Slakter, M.J., Koehler, R.A., & Hampton, S.H. (1970). Grade level, sex, and selected aspects of test-wiseness. *Journal of Educational Measurement, 7*, 119-122.

Smith, J.K. (April, 1980). *The convergence strategy of test-wiseness.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Stanley, J.C. (1971). Reliability. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356-442). Washington, DC: American Council on Education.

Thorndike, E.L. (1951). Reliability. In E.F. Lindquist (Ed.), *Educational measurement* (pp. 560-620). Washington, DC: American Council on Education.

Vernon, P. (1962). The determinants of reading comprehension. *Educational and Psychological Measurement, 22*, 269-286.

The Alberta Journal of Educational Research *Vol. XL, No. 2, June 1994, 213-231*

Bikkar S. Randhawa
University of Saskatchewan

Theory, Research, and Assessment of Mathematical Problem Solving

This article describes theoretical and research issues in mathematical problem solving. Among the theoretical considerations are: nature of problem solving, problem solving approaches, situated and context-independent cognition, externalization of internal representation and its role in the emergence of genius solutions to enduring problems, and the development of competence. The research issues identified concern the assessment of mathematical problem solving of students so that meaningful and contextually relevant analysis of the problem solving process can take place. A study is described in which this approach is exemplified. Also, differences in cognitive and metacognitive processes of high and low mathematics ability students in solving five everyday mathematics problems are related to the literature on the development of expertise.

Cet article décrit des questions théoriques et des questions de recherche dans le domaine de la résolution de problèmes en mathématiques. Parmi les considérations théoriques on inclus: la nature de la résolution de problèmes, des approches à la résolution de problèmes, la cognition située et indépendante des contextes, l'extériorisation de la représentation interne et son rôle dans l'émergence de solutions géniales aux problèmes persistants, et le développement de la compétence. Les questions de recherche identifiées traitent de l'évaluation en résolution de problèmes en mathématiques des élèves de façon à ce que l'analyse des processus de la résolution de problèmes soit significative et vue dans son propre contexte. Une étude illustrant cette approche est présentée. En plus, les différences entre les processus cognitifs et métacognitifs des élèves d'habiletés supérieures et inférieures en mathématiques à résoudre cinq problèmes mathématiques ordinaires sont interprétées selon la littérature dans le développement de l'expertise à la résolution de problèmes.

This article deals with theoretical and research issues in mathematical problem solving. Theoretical issues concern the nature of problem solving, approaches to problem solving, situated and context-independent cognition, externalization of internal representation, and the development of competence. The research and assessment section deals with one of our research studies in mathematical problem solving and gender differences, not a primary focus here.

Problem solving in general and in specific mathematics domains have long occupied researchers. This area received more impetus with the reemergence of cognitive psychology in the 1950s. It is not surprising that this interest coincided with an increased interest in including critical and creative thinking curricula in grade schools and postsecondary institutions. Often critical and creative thinking was seen as a new discovery. Its advocates began promoting the development of these abilities as if past generations had grown up without

---

Bikkar Randhawa is a professor of educational psychology with current research interests in assessment, cognitive psychology, and gender-related differences in learning, instruction, and problem solving in mathematics.

them. Ironically, those promoting these abilities were themselves schooled where these abilities were not emphasized. The question arises how and why those who began to see the need to develop problem solving competence came to this realization. Other related questions are: How do people acquire problem solving competence? Are those who receive direct or curriculum-infused instruction in problem solving competent in domain-specific contexts or in general situations? In other words, do these abilities transfer in nonsituated applications?

Mathematics is a subject in which problem solving is most critical (Polya, 1962). Despite the importance of problem solving ability, however, little is known about how to teach people to solve problems. A great mystery of our time is why some people solve problems for which they have all the necessary component skills and others fail.

Polya's (1962) contribution in the domain of mathematical problem solving received general acclaim (Scandura, Durnin, & Wulfeck, 1977). Also, there was intrinsic support for his notion of heuristics. "Although often useful, his heuristics are frequently little more than general hints, and leave much to be desired insofar as pinpointing what a human or computer must know in order to solve specific kinds of problems" (1977, p. 166). Theoretical and research issues in mathematical problem solving, as evidenced above, are complex. Complexity does not mean we should ignore or not debate these issues. This article provides a brief discussion of the issues identified above.

Theoretical Issues

The Nature of Problem Solving

Problem solving has meant different things to different people and has given rise to different interpretations in the same or different domains of study. Problem solving is not restricted to mathematics. Creating new ideas, for example, is a form of problem solving. Troubleshooting and inventing new products or techniques are other forms of problem solving. Branca (1980) states:

> Although problem solving in mathematics is more specific, it is still open to different interpretations. Activities classified as problem solving in mathematics include solving simple word problems that appear in standard textbooks, solving nonroutine problems or puzzles, applying mathematics to problems of the "real" world, and creating and testing mathematical conjectures that may lead to new fields of study. (p. 3)

Problem solving has meant (a) a goal, (b) a process, and (c) a basic skill. This characterization does not include the role of knowledge. Greeno (1980) emphasized the important role of knowledge in problem solving. Knowledge of a discipline is important in problem solving in that discipline and that is precisely what we know from research on expertise and problem solving. Greeno described the distinction that used to be made between performance based on knowledge, that is, applying an algorithm or just following a remembered set of sequences, and performance based on problem solving. The latter was judged more highly than the former. Here the value on performance was placed by the judge. If the performance was such that the judge could identify the representative knowledge used in solving the problem then the perfor-

mance was judged less highly than if the knowledge base for the performance was not identifiable by the judge. Because of the research in problem solving this distinction is much eroded. But folklore still prevails. Greeno (1980) argues:

> All problem solving is based on knowledge. A person may not have learned exactly what to do in a specific situation, but whatever the person can do requires some knowledge, even if that knowledge may be in the form of general strategies for analyzing situations and attempting solutions.
> The other reason that the distinction between knowledge-based performance and problem solving has eroded is that we now can characterize the performance of individuals who solve problems; and when we carefully consider the performance that occurs in more routine situations, we find that the essential characteristics of real problem solving are there also. (p. 10)

Often major components of problem solving are present in situations where we do not regard the performance as an instance of problem solving. So it is important to distinguish between: (a) situations where the problem solver has specific domain knowledge that considerably facilitates problem solving; and (b) situations where the problem solver must use more general knowledge and procedures to solve a given problem. But the specificity of relevant knowledge is a matter of degree, not kind (Greeno, 1980). Problem solving performance should be considered a continuum, not a dichotomy.

The value we place on problem solving performance has implications for both instruction and assessment. If we do not value the role of specific domain knowledge in problem solving, it implies that instruction in the specific underlying knowledge would not be necessary. Therefore, instruction will emphasize only generic problem solving skills and strategies so that performance will merit high value when it is judged. Proportional weights for domain-specific knowledge and generic problem solving skills and strategies would be required in order to undertake meaningful assessment in this area.

Approaches to Problem Solving: Representational and Strategic Processes

Mathematical problem solving task can be broken down into two major component processes: (a) problem comprehension, and (b) problem solution (Mayer, 1985, 1986; Mayer, Larkin, & Kadane, 1984). Problem comprehension involves translation of each sentence in the problem into an internal representation and integration of the literal information to form a coherent structure. Problem solution, on the other hand, requires planning, monitoring, and executing the requisite mathematical operations. Gagne (1983), however, suggests that there are three phases in the performance of a mathematical task: translation, execution (problem solution), and validation or monitoring of the problem solution. It has been found that students usually have trouble in the problem comprehension or translation phase (Mayer, 1985, 1986). However, most mathematics instruction focuses on the problem solution phase, particularly algorithmic computations (Carpenter, Corbitt, Kepner, Lindquist, & Reys, 1981; Randhawa & Beamer, 1990). Also, students put little emphasis on the monitoring phase.

In somewhat the same vein Kintsch and Greeno (1985) presented a processing model that dealt explicitly with both the text comprehension and problem

solving aspects of arithmetic word problems. The main features of this model include a set of knowledge structures and strategies for using them in building a representation and solving a problem. The verbal input is transformed into a conceptual representation of its meaning that takes the form of a list of propositions. The knowledge structures of this model comprise (a) a set of propositional frames, (b) a set of schemata that represent properties and relations of sets in general form, and (c) a set of schemata that represent counting and arithmetic operations in general form.

Kintsch and Greeno (1985) presented evidence to suggest that the general features of the comprehension process are alike in a variety of situations such as mathematical word problems, reading a story, and reading a stock market bulletin, but that the content of the comprehension strategies, the nature of the specific knowledge structures, and the type of resulting macro structures are task- and goal-specific. In their view, comprehension of a mathematical word problem is achieved by constructing a conceptual schema from the verbal form of the problem on which problem solving processes can operate. For example, understanding a mathematical word problem would constitute conceptual relations among numerical values that guide the selection of calculations to be done. Students' experience in solving word problems "results in their acquiring a special set of strategies for constructing mental representations of texts that are suitable for applying mathematical operations" (p. 110). In this sense, understanding and solving word problems can be regarded as "expert" performance as cited in the literature on problem solving in physics (Chi, Feltovich, & Glaser, 1981; Larkin, McDermott, Simon, & Simon, 1980).

Sweller, Mawer, and Ward (1983) found that mathematics experts (graduates) and novices (9-12-year-olds) employed distinctive strategies while solving mathematical problems. Expert problem solvers preferred forward-chaining strategies, using a large formula, as opposed to various steps one at a time, whereas novices typically employed a means-ends approach, working backward from the target solution. The novices' use of means-ends strategies delayed acquisition of an appropriate schema. Also, the use of nonspecific rather than specific goals was found to facilitate acquisition of expertise. The course of development of expertise witnessed the switch from the means-ends strategy to the forward-chaining strategy. Also, as expertise was acquired, subjects showed the conventional concomitants of expertise such as a reduction in the number of moves required for solution. These results and their conclusion flout conventional wisdom in presenting problem structure. For instance, they conclude

> that the real goal of problem solution needs to be analyzed. If the goal is simply to solve a particular problem, then a conventional problem format is satisfactory. Alternatively, a reduced goal-specificity procedure might be preferred if, as normally occurs in educational contexts, the real goal is to gain additional insights into the problem structure and the structure of the subject matter that gave rise to the problem. (p. 661)

Hall, Kibler, Wenger, and Truxaw (1989) analyzed the episodic structure of written protocols of 85 undergraduates who were asked to solve four story problems. Their results showed that comprehension and solution of the problems were complementary activities, rather than distinct phases of a prob-

lem solving task as Gagne (1983) and Mayer (1985, 1986) proposed. Also, Hall et al. (1989) found that the two complementary activities in the problem solving process resulted in a succession of episodes. Furthermore, competent problem solvers used various forms of model-based reasoning to identify, pursue, and verify quantitative constraints required for the solution. Russell and Ginsberg (1984) reported that fourth graders with mathematical difficulties or novices lacked sophisticated understanding and strategies, but they were not seriously deficient in basic mathematical concepts and nonalgorithmic procedures.

Problem solvers' earlier experiences with a genre of problems are linked to their self-confidence, a crucial factor in problem solving (Lester, 1980). Self-efficacy, confidence in one's ability to do a task rather than actually doing it, is important for cognitive functioning (Bandura, 1986; Davis, 1973). Furthermore, Bandura (1986) and Polya (1957) emphasized the role of the problem solver's interest and motivation. They also stressed that those presenting the problem for solution must choose an appropriate level of difficulty and present it in an interesting way.

Snyder (1987) identified two types of people in social contexts, high self-monitoring and low self-monitoring. High self-monitors tend to be sensitive to feedback and use this to provide cues to guide their behavior. Low self-monitors are not as responsive to the environmental cues. Wong (1988) tested this theoretical formulation in mathematics problem solving and found that high self-monitoring individuals were significantly better at applying rules and in general problem solving abilities.

Lewis and Mayer (1987), in a study involving two experiments, investigated college-level students' difficulties in comprehending relational statements in arithmetic word problems. Their results supported the contention that students' errors on many mathematical problems result from difficulty in the comprehension phase rather than in the solution phase. The proposed theoretical model states that "the problem solver comes to the problem solving task with a set of schemata or preferences for the form of assignment and relation statements in compare problems" (p. 370). When the form of the relational sentence is incongruent with the problem solver's schema, he or she must restructure it, and this rearrangement or transformation may lead to representational errors. This formulation for the specific type of mathematical problem can be deduced from the general schema preference or bias of the problem solver. For problems that are not similar in structure, the general expectation for each is derived from earlier experience (Bandura, 1986).

Situated and Context-Independent Cognition

People bring their personal schemata to problem solving tasks. These personal schemata, knowledge structure, strategies, and preferences, guide how people approach problem solving. Critical to developing problem solving schemata is experience with the genre of problems (Kintsch & Greeno, 1985; Sweller et al., 1983). It is reported that boys have more access to problem solving experiences outside the mathematics classroom than do girls (Kimball, 1989; Hyde, Fennema, & Lamon, 1990), and it is possible this experiential advantage would translate into differential expertise of boys and girls on similar tasks or tasks that require situated learning (Brown, Collins, & Duguid, 1989; Lave & Wenger, 1991). Because of this experiential advantage outside the classroom,

even when boys and girls are matched for academic achievement in mathematics, boys' cognitive and affective schemas might be more appropriate for handling everyday mathematical word problems.

Knowledge is situated because it is in part a product of the activity, context, and culture where it is developed and used. This view of knowledge affects our understanding of learning and problem solving. Brown et al. (1989) argue that conventional schooling too often ignores the influence of school culture on learning and problem solving. The Cognition and Technology Group at Vanderbilt (1990) propose anchored instruction to overcome the problem of *inert knowledge* (Whitehead, 1929), which is knowledge that can usually be recalled when individuals are explicitly asked to do so but is not used spontaneously for problem solving even when relevant. Situated cognition and anchored instruction, which provides a way to recreate some advantages of apprenticeship training in formal educational settings involving groups of students, suggest ways to think differently about instruction for and assessment of problem solving.

Externalization of Internal Representation

Problem solving in any domain requires, among other things, encoding and translation, which are internal and unobservable. To understand the process of problem solving these internal mechanisms must be externalized. The form of these representations is not always verbal; sometimes internal representation consists of internal perceptual images. Visualization as a strategy in problem solving and skillful performance is assumed to play an important role in many problem solving situations (Paivio, 1978; Shepard, 1978). Whether verbal or nonverbal, externalization of internal images offers a richer data base for understanding problem solving processes. Shepard (1978), in particular, depicts a large array of his externalization of internal images. He also documents the use of imagery by many well-known creative geniuses including Albert Einstein, Hermann von Helmholtz, Sir Francis Galton, James Clerk Maxwell, and James D. Watson.

As shown above, visualization was important for many important discoveries. Therefore, it must have implications for mathematical problem solving in a variety of contexts. But problem solving is discovering new applications of acquired knowledge (Greeno, 1980), and "knowledge itself comes to form a part of a person's context of problem solving" (Sternberg & Salter, 1982, p. 19). The challenge is to seek ways to assess the visualization. Assessment of visualization is as challenging as assessment of other internal thought processes engaged in by the problem solver during the process of problem solving. In this context it is imperative to consider the development of competence in problem solving.

Development of Problem-Solving Competence

Competence is the capacity to perform to a set of criteria for a job, task, or ability. It underlies the accumulation, through experience and training, of the required domain knowledge base to function adequately when given an opportunity to perform. The observable expression of competence—performance—is viewed in the literature as depending in part on the context in which performance occurs (Zigler & Seitz, 1982).

Competence in problem solving develops with experience in a situated context (Brown et al., 1989; Lave & Wenger, 1991). Problem solving competence in a cognitive domain is much like a skill in a particular domain in that its transfer is limited. Zigler and Seitz (1982) give an example of a handicapped individual who could operate a complicated piece of machinery competently; but when the machine was rotated 30 degrees, the individual was confused and unable to operate it. One could argue that this person was less competent than a skilled machine operator; more to the point, however, is the fact that, in a well circumscribed context, he or she was able to demonstrate skilled performance. What debilitated this individual's performance was its unfortunate lack of cross-situational transfer.

Sternberg and Salter (1982) point out important assessment-related issues in this area. They state:

> The sensitivity of performance to context suggests that methods to assess competence—that is, ways of attempting to elicit performance—must be designed with consideration of contextual factors, and must have considerable diversity to reflect the diversity of possible contexts in which performance may occur. This observation obviously has implications for testing and measurement. It also seems to require that a full theory of intelligence address the ways in which various contexts may affect the expression of a given competence. (p. 18)

Because developing problem solving competence depends on experience in a specific context, the level of competence at various ages would vary. Thus the range of individual differences in the expression of competence in mathematical problem solving should parallel that in other cognitive abilities. This experiential perspective of development of competence in mathematical problem solving does not exclude genetic factors; rather, it regards these sources of individual differences as putting an upper limit to the extent to which competence can be developed.

Research and Assessment of Mathematical Problem Solving

This section describes a research study in which assessment of mathematical problem solving was undertaken. This example should underline an approach to authentic assessment of problem solving in which multisource data were recorded for analysis and evaluation.

Introduction

Knowledge, transformational operations, and concomitant skills or models are necessary but not sufficient for skilled performance. At times people do not perform optimally although they may possess all the necessary knowledge structures and strategies. This is because "self-referent thought mediates the relationship between knowledge and action" (Bandura, 1986, p. 390). The representation of the past experiences involves constructive processes, which personalize the representation (Bandura, 1986). Moreover, environmental events are filtered through personal meanings and biases and are cognitively transformed into propositional knowledge. When subjective environmental experiences are not congruent with belief, memory for such past experiences is distorted (Cordua, McGraw, & Drabman, 1979; Signorella & Liben, 1984).

The purpose of this study was to assess the cognitive and metacognitive skills, problem solving achievement, and related processes of high and low mathematics ability students on mathematics problems.

Method

Subjects

Forty subjects who participated in this study were selected from a population of 225 grade 12 students in nine classes from three high schools in a medium-sized city in midwestern Canada. All 225 students were enrolled in Algebra 30, an academic grade 12 algebra course, and all had taken the Mathematics Ability Test (MAT) at the beginning of their Algebra 30 course. The MAT is a 40-item test representing the content of academic grade 11 mathematics (Randhawa, Ng, & Beamer, 1990).

Twenty students, 10 boys and 10 girls, were selected from the top quartile of the MAT test (a high ability or HA group); and another 20 students, also 10 boys and 10 girls, were selected from the bottom quartile of the MAT test (a low-ability or LA group). Students were matched pairwise by gender.

Instruments

Development of everyday mathematical problems. Twenty-eight mathematical word problems reflecting real-world, everyday situations were developed. These problems were considered appropriate for those students who had finished the grade 11 academic mathematics course. The problems covered a variety of mathematical content areas including area, proportion, ratio, graphing, tables, charts, percentages, multiplication, and division. The situational content of the problems entailed, for example, car rental, sale prices, currency exchange, and discounts.

Two grade 12 teachers were enlisted, along with two of the researchers, to rate the 28 problems on their relevance to everyday life and their level of difficulty. Because of the ratings only 24 problems were retained for use in the study. Three problems rated as easy were used as warm-up problems, and every student attempted one of these at the beginning of the first videotaping session. Five were selected as the core problems and every student was required to solve these. The problems appear in Appendix A.

The remaining 16 problems were designated as optional and were administered to the subjects only to get provisional data for assessing their effectiveness. With three exceptions, data for only eight subjects for each of these optional problems could be collected under the time constraints of the study. For three problems data were collected on fewer than eight subjects. Therefore, other details on the classification and administration of the optional problems are not relevant for the purposes of this study.

Each problem was prepared on a single sheet, which allowed for work space for students to use during their problem solving session. Additional paper was provided as needed. Some problems had documentary attachments on separate sheets.

Procedure

Each student attended two sessions. In the first each student was given nine problems to solve. In order of presentation, each student was given a practice problem, an optional problem, five core problems in the same order to each

subject, and two of the optional problems. The students were asked to think aloud while solving the problems as it was a purpose of the study to find each subject's representation of the problem as clearly as possible. The subject worked on the solution of the problem uninterrupted by the experimenter unless the subject stopped verbalizing the procedures and strategies being executed. In that case the experimenter would encourage the subject to reveal his or her mental operations by verbalizing.

Each session was videotaped. For this purpose an Elmo Visual Presenter, EV-308, was interfaced with a VCR and a monitor. Through a camera focused on the problem sheet, this equipment can capture the written work, and through a microphone the oral record of the subject; confidentiality was maintained as the subjects themselves were not pictured.

In the second session each subject was interviewed and questioned on the way he or she had represented each problem, the strategies used, and the various procedural steps taken to solve the problem. During this interview the subject was shown his or her work on the problem sheet from the first session. The equipment used in the first session was again used in the second session for a written and oral record of the interview.

One of the two researchers interviewed the subject. The other researcher was present at all the sessions to operate the videotape machine and to take additional notes when required.

After each subject had been interviewed on the specifics of each problem, he or she was asked to rate the problem on its familiarity (how familiar he or she was with the specific kind of problem) and everyday occurrence (what was the perception of the subject about how often a specific type of problem was encountered by people in their everyday business) using a 5-point scale. Only the end points of the familiarity scale were described: 1 = "not at all familiar"; and 5 = "extremely familiar." Descriptors for the everyday occurrence scale were: 1 = "never"; 2 = "once or twice a year"; 3 = "about once a month"; 4 = "about once a week"; and 5 = "everyday."

Each session took about an hour. The interview timetable for each student was arranged such that the second session was on the school day after the first session. The subjects were paid for their participation in the experiment.

Analysis

Problem Solving Skills

Videotaped segments of each respondent, while solving the five core problems and while being interviewed about the same five problems, were analyzed by two trained coders. The coders used a five-point Likert scale to rate the level of evidence of nine cognitive and metacognitive skills of each subject. The five points of this scale were: 1 = "no evidence"; 2 = "below average evidence"; 3 = "average amount of evidence"; 4 = "above average evidence"; and 5 = "substantial evidence." For each subject the ratings of each skill evaluated by the two coders were combined. The consistency between the two coders for rating each skill was determined by computing the Pearson correlation coefficient.

The nine cognitive and metacognitive skills rated were: declarative knowledge; procedural knowledge; translation of mathematical information; use of heuristics; use of estimation or checking or monitoring the solution; level of confidence; ability to transfer the current knowledge; ability to general-

ize/abstract; and mental computing facility. These categories are often mentioned in the literature as important problem solving skills (Bandura, 1986; Carpenter et al., 1981; Chi et al., 1981; Mayer, 1986), but assessment of these skills using rating scales was adapted from a recent study (Randhawa & Beamer, 1990).

The sum of the ratings of the two coders was analyzed in a 2 x 2 (Ability Level x Gender) fixed factorial analysis of variance (ANOVA) design for each cognitive and metacognitive skill. Students' ratings of their familiarity with each problem and of their perception of the occurrence of the kind of problem in everyday transactions were analyzed for each problem in the same ANOVA design.

Accuracy and Monitoring

Significance of differences of percentages of students who monitored their work/solution for each problem for the two ability levels and gender categories was tested.

Each core problem was scored by two independent scorers according to a predetermined scoring scheme. This scheme took into account both the process and computational steps completed for each problem. Depending on this analysis, the total score for a problem ranged from 4 to 17. Score assignments for the process and computational components of each problem are provided in Appendix A.

Using the 2 x 2 ANOVA design described above the total score on the five core problems and the scores on each problem were analyzed. Despite the repeated measurement of the five problems, it was felt that analysis of each individual problem would provide more insight into the differences for the competence level and gender. A further justification is that each problem requires different strategies and representational processes for solution.

Table 1
Mean of Combined Ratings of Cognitive and Metacognitive Skills by Ability and Gender

	Ability		*Gender*		MS_e
Problem Solving Skill	*High*	*Low*	*Male*	*Female*	
Cognitive					
Declarative Knowledge	7.75*	4.35	6.50	5.60	3.68
Procedural Knowledge	8.25*	4.85	7.15*	5.95	2.58
Translation of Information	7.45*	4.35	6.70*	5.10	2.47
Ability to Transfer Current Knowledge	7.35*	4.40	6.20	5.55	2.30
Ability to Generalize/Abstract	6.95*	4.65	6.05	5.55	2.26
Computing Facility (Mental)	7.75*	5.35	5.70*	6.40	1.03
Metacognitive					
Heuristics	7.40*	4.65	6.50	5.55	2.90
Estimating/Checking	7.30*	4.75	6.50	5.55	2.86
Level of Confidence	7.50*	4.20	6.60*	5.10	2.54

*$p<0.05$.

Results

Cognitive and Metacognitive Skills

Consistency estimates of the two raters' analysis of various cognitive and metacognitive skills of students were in the range .904 to .956.

Means of combined ratings of cognitive and metacognitive skills by ability and gender are provided in Table 1. Also shown in this table are the MSe for the univariate effects. The profile of means of the HA group was significantly different from that of the LA group, multivariate $F(9, 28)=7.06$, $p<.05$. The ANOVA *F*-ratios with 1 and 36 degrees of freedom indicated that the ability main effects were significant ($p<.05$) for all the skills investigated. In each case, as expected, the high achiever group (HA) were rated significantly higher than the low achiever group (LA).

The multivariate gender main effect was significant, $F(9, 28)=3.45$, $p<.05$. But the univariate Gender main effects were significant ($p<.05$) only for the Procedural Knowledge, Translation, Confidence, and Mental Computing. In the Use of Heuristics and for the Ability to Estimate the Answer, for the gender effect, the *F*-ratios only approached significance ($p<.10$). It appears that boys were more skillful in Procedural Knowledge and Translation, and approached mathematical problem solving tasks with more confidence than their female counterparts. Girls, on the other hand, were significantly better than their male counterparts in Mental Computing. On the remaining five skills, the observed means of boys were higher than those of girls, but the differences were not statistically significant. The multivariate Ability x Gender interaction was not significant.

Frequency of Occurrence of the Core Problems

Means of students' ratings of the frequency of occurrence (encounter) in everyday life of the five mathematical problems they solved are given by ability and gender in Table 2. Generally, the HA students indicated more favorably than the LA students that each of the five mathematical problems occurred frequently in daily commerce and life. But significant *F*-ratios ($p< .05$) were obtained only for problems 3 and 4.

Table 2
Means of Student Ratings of the Core Everyday Mathematics Problems for their Frequency of Occurrence by Ability and Gender

	Ability		Gender	
Math Problem	*High*	*Low*	*Male*	*Female*
Core Problem 1	2.30	1.80	2.20	1.90
Core Problem 2	3.65	3.10	3.55	3.20
Core Problem 3	2.40*	1.60	2.25	1.75
Core Problem 4	2.70*	1.65	2.60*	1.75
Core Problem 5	2.45	1.85	2.55*	1.75

*$p< 0.05$.

Table 3
Means of Student Ratings of the Everyday Core Mathematics Problems for their Familiarity by Ability and Gender

Math Problem	Ability		Gender	
	High	Low	Male	Female
Core Problem 1	3.10*	2.35	2.90	2.55
Core Problem 2	4.15	3.65	3.80	4.00
Core Problem 3	3.25*	2.20	2.85	2.60
Core Problem 4	3.35	2.60	3.20	2.75
Core Problem 5	3.50*	2.70	3.15	3.05

*$p < 0.05$.

Similarly, boys generally rated the frequency of occurrence of the five problems higher than did girls, but only for problems 4 and 5 the *F*-ratios were significant at the .05 level. Although these problems were intended to be everyday problems and the judges felt them to be so, the students did not perceive them as everyday mathematical problems.

Familiarity Ratings

Means of students' perceived familiarity with each problem by ability and gender are given in Table 3. Competent students' means of ratings on problems 1, 3, and 5 were significantly higher ($p<.05$) than those of LA students. On the remaining two problems the means of the HA group were higher than those of the LA group, but the differences were not significant.

Table 4
Percentages of Correct Answers on Five Core Mathematics Problems by Ability, Gender, and the Total Group and Standard Errors (s_e) for Testing the Significance of Two Independent Proportions

Problem	Ability Level		Gender		Total	s_e of p
	High (n =20)	Low (n =20)	Male (n = 20)	Female (n = 20)		
1. Part 1	80*	20	65	35	50.0	.158
1. Part 2	45*	5	30	20	25.0	.137
1. Part 3	80*	10	55	35	45.0	.157
1. (combined)	68*	12[a]	50	30	40.0	.155
2.	60	30	55	35	45.0	.157
3.	60*	20	55	25	40.0	.155
4.	50*	15	45	20	32.5	.148
5.	30	15	35	10	22.5	.132
Total	54*	18	48	24	36.0	.151

[a]These are the average percentages obtained by averaging those of the three parts.
*$p < .05$

There were no significant differences between the means of boys and girls on the familiarity ratings of any of the five core problems.

Correct Final Answers

Percentages of respondents getting the correct final answer on each problem by ability and gender are given in Table 4. Also, this table gives percentages of the total group who obtained the correct answer for each problem or subproblem. An examination of these percentages allows a quick conventional analysis of the difficulty of obtaining the correct answer for each problem. As can be seen, in terms of the correct final answer obtained, problem 5 was the most difficult (22.5%) and problem 1.1 was the easiest (50%).

The HA group, as expected, obtained a significantly higher proportion of correct answers on each problem, except problems 2 and 5, than the LA group. There were no significant gender differences in the proportions of correct answers.

Monitoring (Checking) Work/Answers

Table 5 shows the percentages of those who monitored or checked the work during each problem solving session by Ability x Gender and the total group. Differences between the ability and gender categories for the independent proportions on each problem were not statistically significant. This suggests that competence level or gender did not affect the monitoring activity.

Accuracy Scores

Means of accuracy scores for the five core problems and mean square errors are shown in Table 6, which also shows corresponding values for the combined accuracy scores of the five problems. The multivariate effect for testing the differences between the mean vectors of the two ability groups was significant, $F(5, 32)=6.80$, $p<.05$. The ANOVA results showed that the ability main effect was significant ($p<.05$) for all except problem 2. As shown in Table 6, the HA group had higher means on all the problems than the LA group. The same

Table 5
Percentages of Students Who Checked (Monitored) Their Work/Answers by Ability, Gender, and the Total Group

Problem	*Ability Level*				*Total*
	High		*Low*		
	Male (n = 10)	*Female (n = 10)*	*Male (n = 10)*	*Female (n = 10)*	
1. Part 1	20	60	60	20	40.0
1. Part 2	20	10	10	10	12.5
1. Part 3.	30	40	0	20	22.5
2.	10	20	0	30	15.0
3.	20	50	60	80	52.5
4.	40	50	60	30	45.0
5.	20	30	20	40	27.5

Note: There were no significant differences between the Ability and Gender categories.

Table 6
Means of Accuracy Scores on the Core Everyday Mathematics Problems by Ability and Gender

	Ability		Gender		MS_e
Math Problem	High	Low	Male	Female	
1. (9)[a]	5.70*	1.77	4.20	3.27	5.27
2. (8)	5.70	3.95	5.20	4.45	9.83
3. (5)	3.70*	2.20	3.45*	2.45	2.39
4. (4)	1.80*	0.70	1.87	0.62	2.83
5. (17)	12.60*	8.40	10.75	10.25	13.67
Total (43)	29.50*	17.02	25.47	21.05	63.79

[a]These are the possible scores.
*$p < 0.05$.

result, obviously, was obtained for the total accuracy scores on the five problems.

An examination of Table 6 suggests that boys had higher accuracy score means than girls on the five problems, but the multivariate gender main effect for the five accuracy scores on the core problems was not significant. Also, the multivariate Ability x Gender interaction was not significant.

Discussion

The results of this study suggest that competence is characterized by a rich knowledge base and metacognitive strategies. Competent individuals exhibit, much like experts, a confident approach to solving problems (Lester, Garofalo, & Kroll, 1989) that is the result of high perceived self efficacy (Bandura, 1986). Not only did the HA individuals work more accurately and efficiently (Chi et al., 1981; Glaser, 1989), they also found the problems presented familiar to them and perceived problems such as these are to be encountered somewhat frequently in everyday transactions outside the classroom. The procedural skills of the HA subjects were superior to those of the LA individuals, and they also obtained a higher percentage of correct answers than the LA group. It appears that the structure of knowledge of those with high mathematical ability, whom we characterize in this study as HA performers, was pervasive in the solution of the everyday problems (Chi et al., 1981; Feigenbaum, 1989). Also, as suggested by Ericcson and Staszewski (1989), it might be because of superior knowledge structures and cognitive and metacognitive skills that our HA subjects exhibited significantly higher levels of efficiency, judgment, heuristics, and extrapolation in solving mathematical problems than the LA subjects.

From the oral and written protocols of the subjects, it was obvious that the HA group had acquired automaticity in the use of algorithms, and their approach to problem solving was governed by conditions in the problem statement. The LA group, on the other hand, was unsystematic in the use of mathematical operations and at times applied various operations by trial and error to obtain a reasonable answer. Interestingly, however, the LA group, much like the HA group, had an idea of an answer that would be reasonable

but did not know how to arrive at it; that is, they lacked procedural skills (Chi et al., 1981). For example, when working on problem 1.1, which required direct application of a percentage shown in a pie graph to a given population, many students from the LA group multiplied the given number by the percentage figure and obtained a much larger number than the total population. When they realized the answer was unreasonable, they divided the total population by the percentage. After checking the answer they realized again that it was unreasonable. Then, sometimes, they would read the problem again and inspect the pie graph and convert the percentage to its equivalent proportion. Again, trial and error governed their behavior, and even after they had finished exploring options and picked a solution, they were not certain that they had obtained the correct solution.

The total accuracy by problem of the HA group was better than that of the LA group. Thus it may be concluded that the cognitive schemata for mathematical problem solving of HA subjects have the contrasting properties outlined above. Instruction in mathematical problem solving might be more relevant and effective if it included both cognitive and metacognitive processes (Glaser, 1989); if it were recognized that expertise is acquired gradually; and if success experiences were recognized as frequent affective ingredients in mathematical problem solving and in building the new cognitive architecture (McLeod, 1989).

Although it was not the major thrust of this article, unlike much other research on everyday problem solving, this study investigated gender differences in cognitive and metacognitive processes. Consistent with Kimball's (1989) synthesis, our results showed that in spite of equal exposure to formal mathematical instruction and curriculum, differential outside experiences of men and women might produce processing differences. The reported differences in boys and girls who were matched on total mathematics achievement were observed on certain cognitive and metacognitive processes. Girls showed higher mental computing facility than boys, but boys were significantly higher on procedural knowledge, translation of problem information, heuristics, and level of confidence. These results would suggest that men and women have acquired different cognitive structures for processing mathematical problems. Similar differences in processing mathematical instruction in natural classroom environments in grade 4 were reported in a recent study (Randhawa & Beamer, 1990). Taken together, the results of these studies would lead us to speculate that processing differences in boys and girls begin at even earlier age levels. Socialization and sex stereotyping phenomena have been ruled out by some in western cultures as plausible explanations. Then what other factors might be instrumental in producing the observed differences? Perhaps the old interactional resolution to the nature and nurture controversy might be applicable here. That is, biogenetic differences and societal interventions jointly produce enduring differences (Kimball, 1989). Differences ought not be considered as something that must be eradicated, but as something we should capitalize on to use effectively what individuals and groups have to offer. Meanwhile, instructional innovations should be explored to optimize the potential of boys and girls in mathematics problem solving so as to develop a reasonable degree of expertise.

Research on humans and computers seems to suggest that one way to change a novice problem solver into an expert is to provide many examples—including problems that are worked out by the instructor (or book) and problems that are worked out by the student. The examples must help the novice to see the step-by-step procedure, rather than just the final answer (Sweller et al., 1983). Just such an approach, process of problem solving as opposed to the product of problem solving, was first used by Bloom and Broder (1950) to help remedial students succeed in comprehensive examinations at the University of Chicago. Such an approach seems to contradict the idea that general thinking skills can be easily taught; in contrast, it shows that thinking skills and strategies are learned in a particular context such as algebra or geometry and that becoming an expert problem solver takes a lot of specific learning. For example, in our study most high ability students as compared with low ability students used one step calculation in Problem 2 to determine the discounted price of the item in each store, that is, the student would say that at Store A the price will be 20% off, which is 80% and at Store B the sale price is 10% off which has already been reduced by 15%. So Store B would charge 90% of 85%, 76.5%. Thus Store B has a better price. These students demonstrated forward-chaining in such procedural large-step calculations. In contrast, most low ability students would use the calculator to calculate 10% of an assumed price, then would subtract this amount from the original. Similarly, to calculate the price at Store B, they would take several discrete steps, sometimes in a haphazard fashion, and then might or might not eventually compute the correct price. The low ability students thus took longer to find the answer and used means-ends strategy in such situations. Problem 1 similarly posed difficulty for these low ability students, whereas high ability students were generally successful in representing the problem properly and executing the solution by undertaking generally bigger steps toward the solution. Also, the high ability students had more confidence in their answer. The progressive differentiation of students by ability in a domain reflects differences in their development of expertise. Unless remedial and specific training is provided in the domain of interest, marked differences in expertise will emerge. Training would reduce but not eliminate individual and group differences in expertise.

The research example provided here goes beyond understanding differences in problem solving expertise. It illustrates how the problem solving process can be captured and assessed. Not only is it important that think-aloud protocols be rendered as reliably as possible, but that the actual process of problem solution be juxtaposed with it so that analysis and subsequent clinical interview are meaningful and contextually relevant. Our use of the specialized recording device made this possible. For the assessment of cognition it is important to pay special attention to the externalization of problem solvers' internal schemata. Thus videorecording of think-aloud protocol along with the accompanying solution process provides initial data. These data ought to be enriched, as in this study, through clinical interviews to maximally capture the cognitive structure of the problem solver. Having data in such form can provide juncture between descriptive narratives and empirical analyses.

Acknowledgment

This article is based on research supported by the Social Sciences and Humanities Research Council of Canada.

References

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bloom, B.S., & Broder, L.J. (1950). *Problem-solving processes of college students*. Chicago: University of Chicago Press.

Branca, N.A. (1980). Problem solving as a goal, process, and basic skill. In S. Krulik & R.E. Reys (Eds.), *Problem solving in school mathematics* (pp. 2-8). Reston, VA: National Council of Teachers of Mathematics.

Brown, J.S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-42.

Carpenter, T.P., Corbitt, M.K., Kepner, H.S., Lindquist, M.M., & Reys, R.E. (1981). National assessment. In E. Fennema (Ed.), *Mathematics education research: Implications for the 80's* (pp. 22-38). Reston, VA: National Council of Teachers of Mathematics.

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Cognition and Technology Group at Vanderbilt. (1990). Anchored instruction and its relationship to situated cognition. *Educational Researcher, 19*(2), 2-10.

Cordua, G.D., McGraw, K.O., & Drabman, R.S. (1979). Doctor or nurse: Children's perceptions of sex typed occupations. *Child Development, 50*, 590-593.

Davis, G.A. (1973). *Psychology of problem solving: Theory and practice*. New York: Basic Books.

Ericcson, C.W., & Staszewski, J.J. (1989). Skilled memory and expertise: Mechanisms of exceptional performance. In D. Klahar & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 235-267). Hillsdale, NJ: Erlbaum.

Feigenbaum, E.A. (1989). What hath Simon wrought? In D. Klahar & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 163-182). Hillsdale, NJ: Erlbaum.

Gagne, R.M. (1983). Some issues in the psychology of mathematics instruction. *Journal for Research in Mathematics Education, 14*, 7-18.

Glaser, R. (1989). Expertise and learning: How do we think about instructional processes now that we have discovered knowledge structures? In D. Klahar & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 269-282). Hillsdale, NJ: Erlbaum.

Greeno, J.G. (1980). Trends in the theory of knowledge for problem solving. In D.T. Tuma & R. Rief (Eds.), *Problem solving and education: Issues in teaching and research* (pp. 9- 23). Hillsdale, NJ: Erlbaum.

Hall, R., Kibler, D., Wenger, E., & Truxaw, C. (1989). Exploring the episodic structure of algebra story problems. *Cognition and Instruction, 6*, 223-283.

Hyde, S.H., Fennema, E., & Lamon, S.J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139-155.

Kimball, M.M. (1989). A new perspective on women's math achievement. *Psychological Bulletin, 105*, 198-214.

Kintsch. W., & Greeno, J.G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92*, 109-129.

Larkin, J.G., McDermott, J., Simon, D.P., & Simon, H.A. (1980). Models of competence involving physics problems. *Cognitive Science, 4*, 317-345.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York: Cambridge University Press.

Lester, F.K. (1980). Research on mathematical problem solving. In R.J. Shumway (Ed.), *Research in mathematics education* (pp. 286-323). Reston, VA: National Council of Teachers of Mathematics.

Lester, F.K., Garofalo, J., & Kroll, D.L. (1989). Self-confidence, interests, beliefs, and metacognition: Key influences on problem-solving behavior. In D.B. McLeod & V.M. Adams (Eds.), *Affect and mathematical problem solving: A new perspective* (pp. 75-85). New York: Springer-Verlag.

Lewis, A.B., & Mayer, R.E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology, 79*, 363-371.

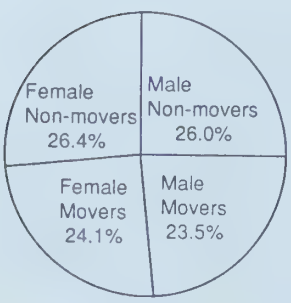
Mayer, R.E. (1985). Mathematical ability. In R.J. Sternberg (Ed.), *Human abilities: An information processing approach* (pp. 127-150). New York: Freeman.

Mayer, R.E. (1986). Mathematics. In R.F. Dillon & R.J. Sternberg (Eds.), *Cognition and instruction* (pp. 127-154). New York: Academic Press.
Mayer, R.E. Larkin, J.H., & Kadane, J.B. (1984). A cognitive analysis of mathematical problem-solving ability. In R.J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (vol. 2, pp. 231-273). Hillsdale, NJ: Erlbaum.
McLeod, D.B. (1989). The role of affect in mathematical problem solving. In D.B. McLeod & V.M. Adams (Eds.), *Affect and mathematical problem solving: A new perspective* (pp. 20-36). New York: Springer-Verlag.
Paivio, A. (1978). On exploring visual knowledge. In B.S. Randhawa & W.E. Coffman (Eds.), *Visual learning, thinking, and communication* (pp. 113-131). New York: Academic Press.
Polya, G. (1957). *How to solve it: A new aspect of mathematical method.* Garden City. NY: Doubleday Anchor Books.
Polya, G. (1962). *Mathematical discovery* (vol. 1). New York: Wiley.
Randhawa, B.S., & Beamer, J. E. (June, 1990). *Learning and instructional processes in grade four classrooms.* Paper presented at the Canadian Society for the Study of Education, Victoria.
Randhawa, B.S., Ng, C.F., & Beamer, J.E. (1990, June). *Mathematical competence: Personal and social influences.* Paper presented at the annual convention of the Canadian Psychological Association, Ottawa.
Russell, R.L., & Ginsberg, H.P. (1984). Cognitive analysis of children's mathematical difficulties. *Cognition and Instruction, 1,* 217-244.
Scandura, J.M., Durnin, J.H., & Wulfeck II, W.H. (1977). Higher-order rule categorization of heuristics for compass and straight-edge constructions in geometry. In J.M. Scandura (Ed.), *Problem solving: A structural/process approach with instructional implications* (pp. 165-199). New York: Academic Press.
Shepard, R.N. (1978). Externalization of mental images and the act of creation. In B.S. Randhawa & W.E. Coffman (Eds.), *Visual learning, thinking, and communication* (pp. 133-189). New York: Academic Press.
Signorella, M.L., & Liben, L.S. (1984). Recall and reconstruction of gender-related pictures: Effects of attitude, task difficulty, and age. *Child Development, 55,* 393-405.
Snyder, M. (1987). *Public appearances/private realities: The psychology of self-monitoring.* New York: Freeman.
Sternberg, R.J., & Salter, W. (1982). Conceptions of intelligence. In R.J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 3-28). Cambridge: Cambridge University Press.
Sweller, J., Mawer, R.F., & Ward, M.R. (1983). Development of expertise in mathematical problem solving. *Journal of Experimental Psychology: General, 112,* 639-661.
Whitehead, A.N. (1929). *The aims of education.* New York: MacMillan.
Wong, N.Y. (1988). Effects of self-monitoring and reinforcement on problem solving performance. *Educational Psychology, 8,* 153-159.
Zigler, E., & Seitz, V. (1982). Social policy and intelligence. In R.J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 586-641). Cambridge: Cambridge University Press.

Appendix A
Core Problems and the Summary of Scoring Scheme for Each Problem

Problems

1. Population 5 years of age and over for Canada, 1981
   Total population 5 years of age and over = 22,280,070

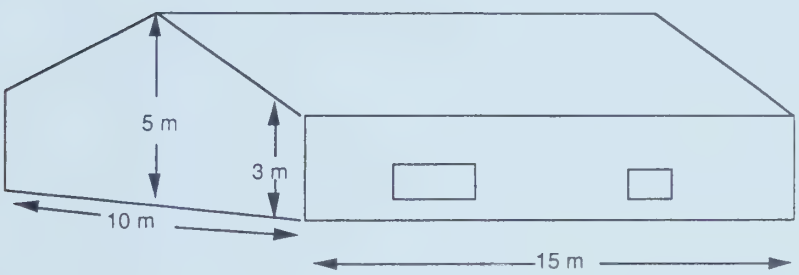


Source: Census statistics, 1981.

Movers are defined as persons who, on Census Day, were living in a different dwelling than the one occupied 5 years earlier.

(a) Calculate the number of female movers.
(b) What proportion of males were movers?
(c) How many more female movers than male movers were there?

2. An identical item is on sale at two stores. At store A, the sale price is 24% off the regular price. At Store B, the sale price is 10% off the price which has already been marked down by 15%. Assuming that the regular prices at both stores are the same, which store has a better offer?
3. A car rental company offers two rental rates for a medium-size compact car. Either $35 daily with unlimited mileage, or $20 daily plus $0.15 per km. Which rate is better to rent a car for a day? Why?
4. Susan paid $600 US for a stereo set when she visited her uncle in San Francisco last week. If the exchange rate was $1 Canadian to $0.85 US, how much did her stereo cost in Canadian dollars?



5. You are going to paint the house shown below with one coat of paint. It has a 1m by 2m picture window at the front and six smaller windows measuring 1m by 1m on the sides and the back. The roof will not be painted.
   A rule of thumb that painters use is one litre of paint for 9.3m^2 of surface.
   The regular price for a 1-litre can is $10.95 but now on sale at 25% discount.
   The regular price for a 4-litre can is $27.95.
   If you cannot use any unused portion of the paint in a can for anything else, what will be the lowest cost to paint the house?

Details of individual score allocations are available from the authors.

Scoring Scheme[a]

Problem	*Process*	*Computations*	*Total*
1.	7	2	9
2.	6	2	8
3.	4	1	5
4.	3	1	4
5.	14	3	17

The Alberta Journal of Educational Research *Vol. XL, No. 2, June 1994, 233-245*

David Bateson
University of British Columbia

Psychometric and Philosophic Problems in "Authentic" Assessment: Performance Tasks and Portfolios

The concept of "authentic" measurement/assessment/evaluation, which has developed and grown from gross misuse and a perception of unfairness and a lack of validity in past measurement, assessment, and evaluation practices, has been rapidly and enthusiastically accepted by many educators. Use of "portfolios" and "performance" tasks has exponentially increased as educators seek to be as "fair" as possible, and to link measurement, assessment, and evaluation more closely to cognition, curriculum, and instruction. However, these "newer" methodologies are fraught with all the problems and pitfalls that have existed throughtout measurement, assessment, and evaluation history. The rigorous employment of high standards of reliability and validity, albeit in possibly different forms from past practice, is equally essential in the use of "authentic" methods and practice as it was in past practice. Much of what has been learned over the long history of non-"authentic" measurement can provide at least the basis for possible solutions to present problems.

Le concepte de mesure/contrôle/évaluation "authentique" qui a pris forme et qui s'est propagé à partir d'un flagrant abus d'emploi, d'une perception d'injustice, et d'un manque de validité dans l'usage antérieur des moyens de mesure, de contrôle, et d'évaluation a été accepté rapidement et avec enthousiasme par beaucoup d'éducateurs et d'éducatrices. Puisque les enseignant(e)s cherchent à être aussi "justes" que possible vis-à-vis l'évaluation, le testing, et la mesure, et de les lier plus intimement à la cognition, au curriculum, et à l'instruction, l'utilisation de "portfolios" et la détermination des habiletés de "performance" s'est accrue de façon exponentielle dans le niveau scolaire. Cependant, ces méthodologies dites "nouvelles" sont chargées de toutes sortes de problèmes et de pièges qui ont toujours existé dans l'histoire de la mesure, du contrôle, et de l'évaluation. L'emploi rigoureux de standards élevés en justesse et en fiabilité, bien qu'il soit utilisé de façons différentes que par le passé, est également essentiel au développement et à l'utilisation des méthodes et des pratiques d'évaluation "authentiques." Beaucoup de ce qui a été appris dans la longue histoire de la mesure non-"authentique" pourrait servir de base à des solutions possibles aux problèmes d'aujourd'hui dans les domaines de l'évaluation, du contrôle, et de la mesure.

During the last three decades, educators (particularly classroom teachers and school administrators) became more and more disillusioned with what they perceived to be inappropriate use of standardized, right-or-wrong, pencil-and-paper, multiple-choice tests. Teachers, administrators, and politicians, in a desire to be as objective (read "fair") as possible, but without adequate knowledge or training about measurement, often grossly misused information from individual tests and testing programs. Although they pointed out that selection-type items, particularly multiple-choice, could not measure many of the

---

David Bateson is an associate professor in the Department of Mathematics and Science Education and a research associate with both the Educational Measurement Research Group and the Centre for Applied Studies in Evaluation. He specializes in science education, measurement, and evaluation.

important goals of education such as the ability to write (as opposed to edit) and the ability to employ complex thinking skills, they continued to depend on such items for lack of any acceptable competitor. The growing negative feelings built up over the years, until the end of the 1980s, when the education community was desperate for change, and a revolution against standardized, selection-type tests was inevitable.

Into the situation came the champions of "authentic" assessment. By its very name, authentic assessment donned a mantle of protective mail. More recently the concepts of "performance tasks" and "portfolios" have also been gathered under authentic's protective wing.

Authentic Assessment

The concept of authentic measurement was first proposed by Archbald and Newmann (1988) and more recently expanded in Berlack et al. (1992). It now dominates discussion of student assessment. A complete shift in basic beliefs and values in measurement has occurred (Kleinsasser, Horsch, & Tastad, 1993) that might be likened to a paradigm shift on the part of many educators, administrators, and politicians.

As an indicator of the extent and effect of this movement, an examination of the program for the annual meeting of the National Council on Measurement in Education held in Washington, DC in 1987, the year before the Archbald and Newmmann document was produced, showed that not one of the 53 session titles contained the words *authentic, performance,* or *portfolio* as they have come to be known in terms of assessment. In 1990, when many psychometricians may still have been hoping that the authentic movement was merely a flash in the pan, there were three papers: two about performance and one about portfolios, but all three dealt with evaluation of teachers rather than assessment of students. By 1993 three of the six presession training courses, and 15 of the regular sessions concentrated on aspects of the authentic movement. A quick glance at the American Educational Research Association programs shows a similar trend.

Change seen in the contents of scholarly discourse reflects the changes in practice in schools. In more and more jurisdictions test scores and grades that in the past parents have thought they understood, are being replaced by anecdotal reports and "narrative profiles" (Moss et al., 1992). The thrust in school reporting has gone from norm-referenced information about the relative ranking of students in various subject areas, to criterion-referenced information about what a given student "can do." Often such "can do" information is given without reference to what a child at that age and/or grade should be expected to do. Faced with this type of reporting, parents complain that they do not understand such information, and they desperately want to know "Where does my child stand?" However, so complete is the transfer to criterion-referenced and can-do reporting, that some schools and teachers are refusing steadfastly to respond to any norm-referenced concerns of the community they serve.

Despite its widespread use and acceptance, it has been difficult to decipher what the term *authentic assessment* means, particularly in relation to current practices. In discussions with university and teacher training personnel, Ministry of Education officials, administrators, and many classroom teachers, it

becomes obvious that there are multiple meanings for the term, depending on the individual questioned and the particular practices employed. However, a common element in most definitions seems to be *any measurement technique that is not a strong component of mainstream, traditional testing and assessment practices up to the date of the coining of the term.* Obviously such a definition or understanding of the term is neither appropriate nor useful in trying to come to grips with what authentic means when used in conjunction with assessment, evaluation, or measurement. From Archbald and Newmann's (1988) original work, authenticity should have two key features:

1. reporting should present the level of mastery of tasks and the nature of the task used to measure that mastery [criterion referencing], and
2. tasks presented to students must be worthwhile, significant and meaningful. (p. 1)

The document goes on to present three principal characteristics of the tasks:

1. production of discourse, things, performances;
2. flexible use of time; and
3. collaboration (pp. 3-4),

and subsequent documents have modified the principal characteristics:

1. collaborations,
2. access to tools and resources,
3. worker discretion or opportunity for ownership, and
4. flexible use of time. (Newmann & Archbald, 1992, pp. 78-79)

Shepherd (1991) makes the best distinction between typical traditional test items and authentic test items: "The [authentic] assessment tasks themselves are real instances of estimators of actual learning goals" (p. 21). Typical traditional test items require the test interpreter to infer from the item or a group of items competence in a specified domain of achievement or interest, whereas authentic tasks are the domain of achievement or interest; interpretation of the item or instrument is not necessary because the assessment is the product itself.

Defining the assessment this way neatly protects it from any questioning about important issues such as are contained in the traditional psychometric views of reliability and validity (e.g., Bateson, 1992). Conversely, any method that is not deemed authentic is, by logic, not authentic, not reliable, not trustworthy, and/or not genuine. One cannot imagine a classroom teacher telling a parent, "Well, we want to do some nonauthentic assessment to alert us to possible problems in reading."

A problem of definition also seems to occur for the related term *performance task*. Unfortunately, it appears that the term *performance* has come to mean, to a large extent, "writing an essay about ..." The National Council on Measurement in Education Presession Course H (Linn, 1991) conducted by the CRESST group, demonstrated scoring rubrics for essays in English, social sciences, mathematics, biology, and physical sciences.

It is difficult, though not impossible, to construe an essay in or about mathematics as being a real or authentic instance of student achievement in mathematics! However, such a trap is relatively easy to fall into because

measurement issues surrounding the scoring of essays have been the subject of development and investigation for many years. The techniques of developing concise, well-defined scales based on specific criteria that are illustrated with exemplars or "anchor papers" representing the range of responses in the total set of responses, and carefully training and retraining of scorers/markers using common illustrative papers, have led to generally acceptable levels of marker agreement in many applications. In the absence of other proven alternative techniques, essay writing has become the primary manifestation of performance assessment.

Despite the comfort of psychometricians with carefully constructed procedures for scoring essays, and the seeming elegance of analyzing and reporting essay results, measurement scholars must not try to misrepresent what is being accomplished through the application of essay writing in the guise of performance tasks: writing an essay is not congruent with doing science just as writing an essay about skiing provides no information about actual skiing abilities. In addition, psychometricians should not become too satisfied with using written expression samples to evaluate students. Despite the increased use of these type of instruments and measures over many years, many of the basic reliability and validity problems have still not been solved. "The development of performance assessments in the early years of the 1980s contained a major set of hurdles that remain to be crossed by those who choose to enter the assessment field" (Chapman & Kerins, 1993, p. 2).

Portfolio Assessment

Theoretical principles have been presented that treat portfolios similarly to writing samples (Reckase, 1993). Four- or five-point scales used to rate the portfolio on various evaluative components are developed, specific written criteria for each scale point are defined and agreed on, and exemplars of each scale point are collected. Raters of portfolios are trained through extensive examination of the scale and each point on the scale, group scoring of common portfolios accompanied by extensive discussion about the judgments made, and teams of scorers working together and sharing ideas.

Despite the theoretical appeal of such procedures and the fact that their application has resulted in high interrater reliability in previous assessments of written expression, studies of large-scale implementations of portfolio assessments have shown that low interrater reliability indices result (Koretz, Stecher, & Deibert, 1993). In the study of the Vermont implementation, the 38 interrater reliability indices for judgments about various aspects and subjects of the portfolios ranged from a low of .28 for the grade 4 "language of math" composite score to a high of .57 for the grade 8 writing "usage" score, and only three of the indices reached .50. The mean of the 38 indices was .37 with a standard deviation of .09 (Koretz, McCaffrey, Klien, Bell, & Stecher, 1993).

One of the major differences between past written expression assessments and modern portfolio assessments is that although the stimulus and response form of the written expression studies has usually been identical for all students, the content of portfolios of work by various students may differ in the extreme. In fact administrators and teachers have been told that "portfolio assessment can take a variety of shapes in a variety of circumstances ... which will permit a range of options, rather than a singular definition" (Fisher, 1991,

p. 10). Although one student's portfolio may contain six examples of writing, a single portfolio may contain a set of pictures of a science fair display, an audiotape of the student reading a narrative passage, and four ribbons from a district track meet. Traditional assessment and testing methods cannot help in developing and validating methods of comparing such diverse data because traditional comparative methods have been based on standardization. Without some standardization evaluating portfolios is as difficult as its parallel problem, the proverbial comparison of apples and oranges.

Although many problems remain to be addressed regarding the implementation of authentic measurement tasks, at least one appears to have been solved, at least for the moment. The standard in terms of methods for developing reliability indices of performance tasks seems to have settled on studies employing generalizability coefficients (Bateson & Blackmore, 1983; Linn, 1991; Trevisan, 1991; Welch, 1993) that were first proposed by Cronbach, Gleser, Nanda, and Rajaratnam (1972), and later developed by Brennan (1983) and Crick and Brennan (1983). Analysis of variance techniques are used to partial the different sources of variance (students, tasks, occasion, scorer/marker, etc., and all the accompanying interactions) and manipulation of the "explained" variables can be used to minimize error.

A Case Example: The BC Science Assessment

The province of British Columbia has joined the states of Connecticut, Vermont, California, and Maryland in radically changing accountability and evaluation systems that have traditionally relied on standardized multiple-choice testing and turning instead to systems that rely more on performance-based tasks and criterion-referenced interpretations (Frechtling, 1991).

With the implementation of the primary program under the auspices of the Year 2000 program (British Columbia Ministry of Education, 1991), the BC Ministry has directed that reporting to parents be essentially interview-type anecdotal records, and have further directed that the use of checklists and standardized norm-referenced tests is inappropriate for primary students. In addition, although the Ministry initially proposed "signposts" and currently has contracts with researchers to develop "reference sets," Ministry expectations are interpreted by primary teachers to mean that even comments based on any criterion-referenced standard are inappropriate; teachers should report only what a student can do without reference to what is normally expected for pupils at that age and grade (British Columbia Ministry of Education, 1991).

The BC Science Assessment, which won the 1993 Division H American Educational Research Association publication award for the best overall instructional program evaluation, was designed specifically to show the benefits of blending traditional and newer forms of assessment practice (Bateson, Erickson, Gaskell, & Wideen, 1992). The overall evaluation used methods and techniques which ranged from strict application of traditional quantitative methods (Bateson, Anderson, et al., 1992); through analyses of students' writing in response to both print and video stimuli regarding socioscientific issues (Gaskell, Fleming, Fountain, & Ojelel, 1992); through description and analyses of closed and free responses to simple science-oriented tasks using materials, and the coding/evaluation of more complex and extensive hands-on investigations (Erickson et al., 1992); to employing videotape, structured observa-

tions, and interviews of administrators, science teachers, and students doing science activities to present the context of science teaching and learning (Wideen et al., 1992).

Although the Ministry of Education usually receives a significant number of complaints from teachers and administrators about the perceived validity and value of each assessment as it is being conducted, for the 1991 Science Assessment, which demonstrated an obvious merging of the many data collection and interpretation methods and techniques discussed above, not only did Ministry officials not receive one single complaint from the field, but they received many laudatory letters. Such acceptance by teachers of external testing was without precedent in British Columbia. For the 1991 BC Science Assessment, teachers perceived the methods and techniques to be appropriate and valid, and many described them as "authentic."

It is essential that more highly visible projects of this nature be conducted. Teachers must see and become part of projects that use a mutually supportive combination of the older, tried-and-true methods and techniques, and the more recently developed methods with which researchers are less secure. It is a time when large-scale measurement, assessment, and evaluation specialists must take some risks. It must be demonstrated, as it was in the 1991 BC Science Assessment, that the best decisions will be made if the data on which the decisions are based come from multiple, complementary methods using a variety of data types. It must be clearly demonstrated that if one relies on a single method or a single source of data, there is a much greater potential for erroneous and invalid decisions. Such demonstrations can have a high impact on teachers and will do much to influence practice, particularly when large numbers of them can be directly involved in the process.

Current Status of Knowledge of Assessment Practice

Many well-thought-out and well-researched techniques for developing "authentic" measures have been published (Shavelson & Baxter, 1992; Swain, 1989). However, these articles/papers have been presented through meetings and journals that are not popular reading for classroom teachers or even administrators. It is essential that messages such as those described above somehow get conveyed to the members of the teaching profession who might be applying such techniques in their classrooms.

The accepted techniques and standards for testing using traditional methods, which have substantially been of a quantitative nature, are well documented in a great number of presentations, articles, and texts. These methods form the foundation not only of the teacher training courses that most teachers have taken in the past, but also of the testing traditions and "folk knowledge" of schools. However, more recently attention has also been directed at qualitative methods of gathering data and research. Despite having entered the mainstream of accepted research methods long after the acceptance of quantitative methods, there has been a rapid acceptance of qualitative research, particularly over the last decade. A major factor in such "softer" methods being accepted has to do with the various challenges to the integrity of the qualitative methodologies that have occurred.

In response to such challenges, researchers promoting the techniques of interviews, observations, and participant observations have developed per-

suasive, logical arguments for their particular procedures and have presented rules of evidence that can be employed to evaluate the procedures. As such, the validity of many models of qualitative research has been established, and these proven qualitative methods and techniques are well suited to use by classroom teachers to provide better data on which to make their educational decisions about students.

The standards for acceptability of both quantitative and qualitative methods, however, are not widely known and understood by most educators, particularly classroom teachers, who are not aware of the many methods and techniques that are available to them. University preservice training has generally failed dismally in preparing teachers for testing, assessment, and evaluation tasks which they must undertake in their classrooms. For the minority of students who actually take a course in testing, measurement, and/or evaluation, the courses almost exclusively deal with the development, evaluation, and use of paper-and-pencil tests to the exclusion of most other measurement and evaluation topics. Even many of the available texts only pay lip service to the topics of alternative assessment procedures.

Without appropriate knowledge and training, teachers have fallen back on the folk knowledge of testing: What was done to them (usually inappropriately). Even the language of psychometrics is usually misunderstood; to most teachers, "validity" is a personal synonym for "good" or "something I like," as opposed to the meanings that test constructors and psychometricians typically ascribe to validity. Most teachers would be amazed if they could understand Messick's (1989) detailed and complex argument concerning validity whose basic message of consequences makes so much common sense. In trying to solve pressing, day-to-day problems of assessment and measurement in their classrooms, teachers and administrators have tried to mimic the techniques and instruments used by large-scale testing and assessment programs that have been the darling of most psychometricians. What they fail to realize is that methods which are generally appropriate for making some types of decisions about groups and systems, or that are designed merely to "red-flag" individuals, are usually not appropriate for making high-stakes classroom decisions about individuals. Teachers and administrators for too long have been the victims and have seen the students for whom they care become victims of inappropriate measurement practices.

In reaction, they have wholeheartedly joined the authentic revolution. Although the established processes and methods of standardized, paper-and-pencil, black and white, selection-type testing have been researched and developed well and have served the purposes of making policy decisions about groups at relatively low cost in the past, the present mushrooming of interest in authentic, performance, and portfolio procedures dictates that the focus of resources, training, and research in the testing/measurement/evaluation community must change.

Maguire (1992) makes the case that authentic assessment must be firmly grounded in educational philosophy and, by implication, cognitive psychology and uses the example of the SOLO taxonomy (Biggs & Collis, 1982) as firmly grounded assessment methodology. Similar calls for firm grounding not only in educational philosophy and cognitive psychology, but also in knowledge of

the structure of the subject matter have been made by other authors (Linn, Baker, & Dunbar, 1991; Moss, 1992). In describing validation criteria, Moss (1992) refers to:

> Content quality: the extent to which the content is consistent with the best current understanding of the field as evidenced by systematic judgments of the quality of the tasks from subject matter specialists and evidence about the ways in which students interpret the content. (p. 248)

Unfortunately, few examples of such sophisticated analysis can be found in classroom, school, district, or system practice. If authentic measures that are being used could point to and provide evidence for such grounding, the issues of validity could more easily be addressed. However, it is usually the case that the derivation or development of such measures is unknown, and the researcher interested in the validity of the measures is left to attempt to infer the underlying educational philosophy and the cognitive psychology bases.

Performance assessment in science is one area where the tasks that have been developed and used to date are well grounded in at least the philosophy of science, and they generally allow students opportunities to demonstrate a range of cognitive complexity. However, few tasks are of quality, and one reason is the extreme difficulty of generating and refining the tasks and the accompanying data collection and scoring/coding criteria. As one group of researchers has described the situation:

> We have found the process of creating performance assessments ... to be time consuming, requiring considerable scientific and technological know-how. Development of quality performance assessments requires multiple iterations through a sequence of development, tryouts with students (getting their thoughts and comments), and revision. Short-circuiting this process leads to ill-conceived and poorly constructed assessments. (Shavelson & Baxter, 1992, p. 231)

The fact that good performance tasks are extremely difficult to develop means that few good performance tasks are available. To compound the problem, because most performance tasks are purportedly designed to assess problem solving and higher level thinking skills, it is essential that the task used for any assessment be novel to the students being assessed. If the students have seen and done the task previously, their performance on the assessment merely becomes a regurgitation of what they remember from the last time they did it. The fact that performance/items must be piloted, often on several occasions, means that the security of the item that is required in order to ensure that the item is measuring what was, or is intended, rather than just recall or replication of previously taught solutions, is compromised.

Teachers are excellent judges of quality performance tasks. As previously identified in the discussion of the definition of authentic measures, because the tasks are designed to be the domain of interest, the tasks are quickly recognized and adopted by teachers as good teaching and learning activities. With the amazing "moccasin telegraph" that exists among teachers, it requires little time for most teachers to become aware of such good tasks. If the item escapes wide distribution during piloting, as soon as it is used to any extent on any

assessment it usually becomes widely known and used by teachers in the course of instruction, and so becomes worthless as an assessment device.

A prime example of such a task is the "paper towels" task that probably originated with the Assessment Performance Unit in Britain, but has now been used in at least 11 jurisdictions on at least four continents and has been the subject of many scholarly presentations, articles, and reports. Teachers who have been involved with any administration, review, scoring, or interpretation tasks associated with the assessments have recognized the elegance of the paper towel task and have used it for teaching purposes as opposed to assessment purposes. Because many students have now interacted with this task, this formerly ideal item/task is now totally useless for assessment of anything but recall of information.

Linn et al. (1991) have proposed eight criteria for judging performance assessments:

1. Intended and unintended consequences,
2. Transfer and generalizability,
3. Fairness,
4. Cognitive complexity of student processes,
5. Meaningfulness of the problems,
6. Content quality,
7. Content comprehensiveness of coverage, and
8. Cost and efficiency. (p. 20)

Transfer and Generalizability Issues

It is difficult to argue against any of the criteria proposed, but merely identifying the general criteria does not define exactly what each criterion means or the evidence that would be acceptable to come to a judgment about it. Without discussing the relative merits of possible ways of implementing the proposed set of criteria as a whole, it is useful to focus on the second criterion, transfer and generalizability. In an expansion of the criteria, Linn et al. (1991) state that the second criterion should be applied as "the degree to which performance on specific tasks transfers" (p. 20). Moss (1992) further expands the meaning of this criterion to include "the extent to which results can be generalized across raters and tasks (reliability, in the traditional sense) and from the specific assessment task to the broader domain of achievement" (p. 248).

With this issue of transfer and generalizability in mind, attempts have been made to standardize performance tasks and to develop generic tasks that purport to measure a generic skill or process without regard to the knowledge and/or interest of the subject matter context (e.g., Jorgensen, 1993). However, such attempts have to date been singularly unsuccessful. Numerous studies of assessments which have attempted to use both subject-independent and structured performance tasks have shown that the context of the performance task is all-important; that generalizable or generic skills and processes that students may possess independent of content, while theoretically appealing, are not possible to operationalize in an assessment situation (Bunch & Littlefair, 1988; Erickson, et al., 1992; Haladyna & Sloat, 1993; Shavelson & Baxter, 1992; Shavelson, Baxter, Pine, & Yura, 1990; Welch, 1993). Even with the scoring of essays, which are the performance tasks which present the fewest problems for psychometricians, "topics are not interchangeable and some students write

better on one prompt while other students write better on other prompts" (Welch, 1993, p. 11). Further, it is also certain that although the writing prompt is identical for all examinees, each examinee's understanding of, and knowledge and interest in, any given task will be somewhat different; "the idea that such an assessment is a fair measure of each respondent's performance is thus open to challenge" (Bunch & Littlefair, 1988, p. 175).

Although it has been possible to develop selection-type testing items, such as multiple-choice, that have high inter-item correlations, and therefore can be inferred to all measure approximately the same construct, similar inter-item correlations have not been found for performance tasks. Tasks which appear to be similar, or appear to require the identical skill applied in the same way but in a different context have shown very low, and often nonsignificant inter-item correlations.

In addition, the nature of performance tasks, particularly those designed to measure the use of higher level thinking and complex problem solving skills, dictates that many of the variables that are fixed in traditional standardized tests are purposely let free to vary in performance tasks. Not the least of the free variables is often the "correct" answer. Whereas in traditional standardized tests there is almost always one single answer that is scored as correct, in performance tasks there may be many more or less appropriate and successful ways to approach the problem. Such being the case, the use of internal consistency measures of reliability will necessarily result in coefficients that are much lower than those obtained from typical standardized multiple-choice tests. For performance tasks, this standard must change—given the nature of performance tasks, internal consistency cannot be expected to be strong and one could argue that internal consistency measures are not even appropriate for use in performance assessments. "Although efficiency, reliability, and comparability are important issues that cannot be ignored in new forms of assessment, they should not be the only, or even the primary criteria to judge the quality and usefulness of an assessment" (Linn et al., 1991, p. 16).

Psychometricians must broaden their typical view of reliability measured through internal consistency to include dependability measures based on inter- (or intra-) rater/observer/scorer analyses. There are also persuasive arguments that even if internal consistency indices are used as the standard of reliability, low indices are worth the price of the information gained.

> The beauty and potential use of the performance assessment is to obtain outcome data on valued achievement targets difficult to assess with standardized achievement tests. The potential gain in further achievement data is obtained at a loss in the magnitude of the reliability coefficient. (Trevisan, 1991, p. 3)

Because most performance tasks, by definition of being authentic measures, actually are the domain of interest, such tasks have great "face validity." Teachers speak highly of them and often will compete to attempt to get their class involved in the assessment. Such competition to be included is diametrically opposite to teacher behaviors in traditional assessment. The perception of such tasks is that they truly measure what should be valued in education. With such wide acceptance by teachers who are traditionally very critical, it is difficult to believe that these performance tasks should be judged to be less

than desirable simply because they are not able to be transferred and do not represent generalizable skills or processes. Because no tasks that have been developed to date can be acceptable based on the proposed criteria of being transferable or generalizable, it may be necessary to reconsider and reconceptualize how performance tasks should be judged.

Future Assessments

Some visions of the future in testing have included alarming scenarios: "NRTs [norm-referenced tests] and CRTs [criterion-referenced tests] have become extinct—as have the spotted owl, African elephant, great white shark, bottle nosed dolphin, and numerous species of birds and insects" (Hansen, 1993, p. 3). If such is to be the case and useful psychometric techniques that to a large extent have been well understood are fated for oblivion, only radical action on the part of psychometricians to demonstrate the benefits of such older methods, similar to radical actions taken by environmentalists, may be necessary. Stiggens (1991) argues that the changes occurring are noteworthy: "Our current assessment upheaval is not simply the latest fad to sweep the education scene. Rather, it signals the end of a 60-year era of educational assessment and our passage into a whole new era" (p. 263).

However, others consider that the present thinking about performance testing and portfolios is just another wave that if merely ignored will recede to the background (Wolmut, 1993).

This type of ostrich-like thinking and action on the part of measurement and evaluation specialists has led to the unfortunate wholesale acceptance of techniques dubbed authentic in today's classrooms. Many psychometricians and others in the measurement community have not responded well to past criticisms, and many are now treating the authentic movement with disdain. They have failed to provide widespread, useful guidance for classroom practitioners in the past, and have more recently failed to respond appropriately to the rapid acceptance by classroom teachers of an enticing solution to what these teachers have seen as the testing monster.

Rather than joining the rush in order to ensure that the development of the authentic movement proceeds in an orderly fashion, many knowledgeable measurement scholars have either ignored what is happening at the classroom level, or have steadfastly but vainly decried what they see as a threat to their comfortable position. They have vigorously attacked the whole concept of authenticity. Although they may have excellent grounds for their attacks, school practitioners are not listening. Fortunately, the tide is turning, and more respected measurement individuals have turned their attention in constructive rather than destructive ways to the changes caused by the ongoing authentic revolution. They are "getting on the bandwagon" rather than trying to "derail the train," and their actions will most assuredly assist in better practices being instituted.

References

Archbald, D.A., & Newmann, F.M. (1988). *Beyond standardized testing: Assessing authentic achievement in the secondary school*. Reston, VA: National Association of Secondary Principals.

Bateson, D.J. (1992). Adding authenticity to sound measurement practice. *Research Forum, 1*(9), 4-8.

Bateson, D.J., Anderson, J., Brigden, S., Day, E., Deeter, B., Eberle, C., Gurney, B., & McConnell, V. (1992). *British Columbia assessment of science 1991 technical report I: Classical component*. Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Bateson, D., & Blackmore, D. (1983, June). *Application of generalizability theory to the number of markers question in scoring of written expression*. Paper presented at the annual meeting of the Canadian Society for Studies in Education, Vancouver, BC.

Bateson, D., Erickson, G., Gaskell, P.J., & Wideen, M. (1992). *British Columbia assessment of science (1991) provincial report*. Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Berlack, H., Newmann, F. M., Adams, E., Archbald, D.A., Burgess, T., Raven, J., & Romberg, T.A. (1992). *Toward a new science of educational testing and assessment*. Albany, NY: State University of New York Press.

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.

British Columbia Ministry of Education. (1989). *Year 2000: A curriculum and assessment framework for the future*. Victoria, BC: Author.

British Columbia Ministry of Education. (1991). *Primary program*. Victoria, BC: Author.

Bunch, M.B., & Littlefair, W. (1988). *Total score reliability in large-scale writing assessment*. Paper presented at the Education Commission of the States, Boulder.

Chapman, C.W., & Kerins, C.T. (1993, April). *How do we report useful information to schools based on authentic assessment*. Paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Atlanta.

Crick, J.E., & Brennan, R.L. (1983). *Manual for GENOVA: A generalized analysis of variance system*. Albany, NY: American College Testing.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurement*. New York: Wiley.

Erickson, G., Bartley, A.W., Blake, L., Carlisle, R.W., Meyer, K., & Stavey, R. (1992). *British Columbia assessment of science (1991) technical report II: Student performance component*. Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Fisher, D. (1991). Portfolio assessment: A framework for a range of options. *Research Forum, 1*(9), 9-12.

Frechtling, J.A. (1991). Performance assessment: Moonstruck or the real thing? *Educational Measurement: Issues and Practice, 10*(4), 23-25.

Gaskell, P.J., Fleming, R., Fountain, R., & Ojelel, A. (1992). *British Columbia assessment of science (1991) Technical report III. Socioscientific issues component*. Victoria, BC: Ministry of Education.

Hansen, J.B. (1993, April). *Assessment in the year 2001: The darkness and the light*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Haladyna, T.M., & Sloat, E. (1993, April). *The influence of context in reporting achievement test scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Jorgensen, M. (1993, April). *Authentic assessment for multiple users*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Atlanta.

Kleinsasser, A., Horsch, E., & Tastad, S. (1993, April). *Walking the talk: Moving from a testing culture to an assessment culture*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program: Interim report*. RAND Institute.

Koretz, D., Stecher, B., & Deibert, E. (1993, April). *The Vermont portfolio assessment program: Interim report on implementation and impact, 1991-92 school year*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Atlanta.

Linn, R.L. (1991, April). *Issues in large scale performance assessment*. Course H training session held at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, Illinois.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Maguire, T. (1992). Grounded authentic assessment and teacher education. In D.Bateson (Ed.) *Classroom testing in Canada: Proceedings of the second Canadian conference on classroom testing.* Vancouver: Centre for Applied Studies in Evaluation.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103) New York: Macmillan.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

Moss, P.A., Beck, J.S., Ebbs, C., Matson, B., Muchmore, J., Steele, D., Taylor, C., & Herter, R. (1992). Portfolios, accountability, and the interpretive approach to validity. *Educational Measurement: Issues and Practice, 11*(3), 12-21.

Newmann, F.M., & Archbald, D.A. (1992). The nature of authentic academic achievement. In A.R. Tom (Ed.), *Toward a new science of educational testing and assessment.* Albany, NY: State University of New York Press.

Reckase, M.D. (1993, April). *Portfolio assessment: A theoretical prediction of measurement properties.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Shavelson, R.J., & Baxter, G.P. (1992). What we've learned about assessing hands-on science. *Educational Leadership, 49*(8), 20-25.

Shavelson, R.J., Baxter, G.P., & Pine, J., & Yura, J. (1990, April). *New technologies for large-scale science assessments: Instruments of educational reform.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Shepard, L. (1991). Interview on assessment issues with Lorie Shepard. *Educational Reseearcher, 20*(2), 21-23.

Stiggens, R.J. (1991). Facing the challenges of a new era of educational assessment. *Applied Measurement in Education, 4(4), 263-273.*

Swain, J.R.L. (1989). The development of a framework for the assessment of process skills in a graded assessment in science project. *International Journal of Science Education, 11*(2), 26-38.

Trevisan, M.S. (1991, April). *Reliability of performance assessments: Let us make sure we account for the errors.* Paper presented at the annual meeting of the National Association of Test Directors, Chicago.

Welch, C.J. (1993, April). *Issues in developing and scoring performance assessments.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Wideen, M., Mackinnon, A., O'Shea, T., Wild, R., Shapson, S., Day, E., Pye, I., Moon, B., Cusack, S., Chin, P., & Pye, K. (1992). *British Columbia assessment of science (1991) technical report IV: Context for science component.* Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Wolmut, P. (1993, April). *Do we know where we're going? Have we been there before? Riding the measurement waves.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

The Alberta Journal of Educational Research *Vol. XL, No. 2, June 1994, 247-262*

Dan G. Bachor

John O. Anderson

John Walsh

and

Walter Muir
University of Victoria

Classroom Assessment and the Relationship to Representativeness, Accuracy, and Consistency

The last decade has witnessed considerable change in the ways teachers assess students in the classroom. The once ubiquitous classroom test has been augmented with a variety of emerging assessment techniques ranging from anecdotal observation to portfolios. Teachers' knowledge of how to evaluate the use of these new classroom-based assessment tools has not always paralleled their introduction. In this article we offer three criteria as guidelines for evaluating these classroom-based assessment techniques, which concern the representativeness, accuracy, and consistency of the information that is collected, interpreted, and reported.

La dernière décennie apporta des changements considérables dans la façon que les enseignant(e)s évaluent leurs élèves en salle de classe. On a ajouté aux tests administrés partout dans toutes les salles de classe, une variété de techniques d'évaluation à partir de l'observation anecdotale jusqu'à l'utilisation de "portfolios" descriptifs. Les connaissances qu'ont les enseignant(e)s à savoir évaluer l'utilité de ces nouveaux outils d'évaluation ne sont pas toujours allées de paire avec l'usage de ces nouveaux outils. Cet article offre trois critères comme guides d'évaluation pour ces techniques d'évaluation en salle de classe et démontre comment l'information collectionnée, interprétée et rapportée est représentative, juste, et constante.

Introduction

In the context of classroom assessment practices, there is an increasing use of a wide array of information collection procedures such as observation and portfolios (Bachor & Anderson, 1993). Although practices vary across grade

---

Dan Bachor is a professor in the Department of Psychological Foundations. His research interests include classroom assessment procedures as well as assessment and instructional strategies for special education.
John Anderson is an associate professor in the Department of Psychological Foundations. His research interests include classroom assessment practices and large-scale assessment.
Walter Muir is a professor in the Department of Psychological Foundations. His research interests, in addition to educational measurement and evaluation, include computer applications to education.
John Walsh is an assistant professor in the Department of Psychological Foundations. His research interests concern self-regulated learning in learning disabled students, cognitive strategy instruction, and assessing student learning.

levels and teachers, for the most part classroom assessment is not a discrete, single-instance group event such as the test. Classroom tests, however, do remain a component of teachers' assessment arsenal, the most common examples being weekly spelling tests for younger children and unit tests for older children. Other more formal tests are rare events and are removed for the most part from classroom activities. When they are given, such measures usually are restricted to a specialized purpose. Examples include administering individualized measures to make placement decisions for students with special needs, or requiring that all grade 12 students take graduation examinations.

Emergent forms of student assessment such as observation and portfolios are considered by teachers to be authentic evaluations of learning outcomes in that they are closely representative of the kinds of ongoing instructional activities in classroom learning (Bachor & Anderson, 1993; Nicholson & Anderson, 1993). Tests, on the other hand, are falling out of favor with teachers because they are seen to be too far removed or decontextualized from ongoing instructional activities (Wilson, 1990). This viewpoint is held by almost all teachers in kindergarten to grade 3 and by a growing number of teachers in grades 4 to 7 (Bachor & Anderson, 1993).

Reporting assessment information also is changing across grade levels. There is a shift from reporting highly condensed symbolic evaluation information in the form of letter grades to more descriptive, textual, or oral reports. Even where letter grades are still reported, they have been supplemented with comments that have tended to become more extensive.

Further, it is apparent that evaluating student achievement remains a common, important, and highly visible component of schooling. One aspect of classroom assessment practice that has received less attention than the particular format of data collection or reporting is that the practices of classroom assessment should themselves be consistently and continuously monitored and evaluated. The purpose of this article is to describe three criteria that can serve to provide a functional and feasible framework for evaluating many of the emergent classroom assessment practices. In meeting this purpose, we do not describe the full range of assessment procedures employed in the classroom. Instead, we restrict our examination to discussing two information collection techniques commonly utilized by teachers of primary (grades 1-3) and intermediate (grades 4-7) children: observation and portfolios (Bachor & Anderson, 1993).

Evaluating Classroom Assessment Practices

Traditionally, the evaluation of student assessment practices has centered on evaluating test quality in terms of validity and reliability. The focus of attention was on one test at a time and on each test administration. In other words, each instance of assessment activity was evaluated. This was generally done in conjunction with formal and often large-scale test administrations. Recently, validity in relation to the evaluation of student achievement has received considerable attention in the educational measurement literature (Messick, 1989; Moss, 1992; Shepard, 1993). In Messick's seminal chapter on validity (1989) and the more accessible chapter of Shepard (1993) the conceptualization of validity has been broadened to include not only the relationship of test items to the underlying traits being tapped, but also to consider the use of the

information generated by the test. Reliability generally has been equated with homogeneity of item performance within the item set composing the test. In addition, reliability is related to consistency of test results given variations in markers, the context of test administration, and in test questions (format and/or topics). The assessment of tests on the basis of validity and reliability would generally be done by measurement specialists associated with the administration and scoring of the test(s) in question.

In contrast, as noted above, classroom assessment practices consist of the melding of different information collection procedures developed and used by teachers that result in a series of data clusters for each student. Any guidelines that might be offered to aid teachers in evaluating their assessment practices must attend to the growing complexity of contemporary classroom assessment and be sufficiently straightforward to be worthwhile to classroom teachers.

Evaluation assessment practices can take the form of an appraisal of the discrepancy between intention and practice: what was intended to be assessed versus what was actually assessed. In the classroom context at least, there also must be a consideration of the quality and extent to which intention has been determined and articulated. That is, the intention must be clearly stated in terms of the established goals of instruction as translated into classroom practice. At least three criteria could be used to guide this evaluation of assessment practice. First, the extent to which assessment practices are based on tasks and activities that are representative of instructional intentions must be determined and then translated into specific information collection, interpretation, evaluation, and reporting procedures. Second, the accuracy of the assessment practices must be addressed. Third, the issue of consistency of these practices over time and across students warrants attention. A basic assumption underlying this discussion of the evaluation of assessment practice is that the classroom instruction is goal-directed and that these goals can be clearly known and described. Further, students will be assessed on the basis of the completion of some task such as writing an essay, responding to a multiple-choice test item, conducting a laboratory exercise, or demonstrating a physical skill such as dribbling a basketball. Finally, it must be noted that any change in the purpose of assessment should result in a revisitation of the specific application of the criteria of representativeness, accuracy, and consistency to any assessment task.

Representativeness

Assessment practice is representative to the extent that the task the student is to complete is clearly derived in a meaningful sense from the intended learning (the instructional) goals. For example, consider the goal of assessing problem solving. If the task may be successfully completed through the recall of information (such as remembering 5+6=11) as opposed to thinking critically about it, then the task would not be representative. A task that caused students to engage intentionally in critical thinking would be representative: 5+6 would be representative of problem solving if students did not know the answer. Instead, they reached the answer by problem solving by noting that 5+5=10 and that 6 is one more than 5, therefore 15+6=11.

Representativeness, then, refers to the extent to which the assessment information characterizes the learning agenda under consideration, where the

learning agenda includes both the broadly stated goals and expectations for students at a particular age-grade level and the more specific instructional targets set in any classroom. Wiggins (1989) may have had representation in mind in describing the concept of authentic assessment as an attempt to shift the focus of assessment practice toward collecting information on student performance that is more directly representative of the kinds of achievement that are valued.

In establishing a representative task in an assessment sense the learning goal has to be identified and described. The description should include as components both the process and product aspects of student achievement of the learning goal. In regard to process, attention should be given to the cognitive processes students use in the assessment task or activity. To illustrate, if multiple-choice questions were selected as an assessment device, the primary focus probably would be on the obtained answer, the product. An assessor might be able to infer how an answer was reached if there were a sufficient number of examples of any one type of answer. When observation is selected as an assessment tool, the tendency is to shift more to noting process.

Finally, when think-aloud protocols are selected, the goal is to unfold how someone is thinking. Thus with a shift in focus the emphasis in assessment can change from either product to process or from process to product.

The first step in selecting a representative task is the identification of valued learning outcomes. These are described in a manner that relates to observable activities or products and that captures the intent regardless of how complex or latent the outcome or goal. In the British Columbia context a major source of general descriptions of learning goals are Ministry program or curriculum documents (BC Ministry of Education, 1990a; 1990b; 1992). Another step in constructing or describing a representative set of activities is related more directly to assessment practices, that is, the development of evaluation procedures that capture both the process and the outcome.

Two examples are illustrative of these steps. In the first the initial task would be identifying and describing a learning outcome such as *communication of direct experience* in a manner that captures the intent and conveys the goal in an instructionally meaningful manner. This description then would be representative and would lead to identification of relevant indicators of achievement. The next step could involve asking students to compose an account of an experience they have had. This probably would be a more representative means of assessing writing ability than using a more indirect measurement format, such as giving a series of multiple-choice questions to check any student's understanding of the requirements of composition or determining if the child knows how to construct a sentence. As a second example, establishing an observation schedule that would result in samples of student behavior taken over time and across types of learning outcome is more representative of the full range of each student's repertoire of behaviors than completing a behavior checklist that calls for a composite picture based on recollections of past school-related activities.

Another aspect of developing representative assessment tasks concerns the extent to which such tasks reveal the cognitive activities of students. Assessment should provide opportunities for students to leave evidence of the

strategies and thoughts that they use of produce assessment products or other achievement-related activities. Just as the products and activities must be reflective of the content objectives of the curriculum, so too must assessment represent the cognitive strategy objectives that form the broad, but important goals of schooling. It is also important to recognize that these cognitive strategies will be utilized in situations that call for problem solving in an affective-laden situation (e.g., how do I avoid fighting when I am angry?) or when psychomotor skills are needed (e.g., how can I use imagery to relax before playing an important football game?) as well as in more obvious cognitive activities such as learning mathematics.

Developing assessment techniques that validly represent student cognition in addition to the products of such cognition is most difficult in the applied setting of the classroom (Marx, Winne, & Walsh, 1985; Royer, Cisero, & Carlo, 1993; Weinstein & Meyer, 1991). Certainly the precision of the cognitive psychology lab cannot be expected. However, there is emerging evidence that students can be trained to produce veridical information about cognitive strategies or plans that they use to accomplish classroom products.

A good example of this form of assessment is found in Swing and Peterson's (1988) study of mathematics problem solving. In this investigation students were trained to solve problems using adjunct questions to guide their problem solving tactics. Swing and Peterson assessed students' strategic use of these prompts by asking them to draw lines linking adjunct questions to appropriate parts of the problem. Of particular importance for the purposes of our discussion is that the students in this example left evidence of their strategic thinking as they solved problems. In this way, Swing and Peterson assessed both the products of student learning, that is, the completed problem solution, and the strategic thinking that produced the products, that is, students' use of adjunct questions.

The assessment of student cognition is not limited to the cognitive strategies of plans that students employ when responding to classroom work. Students can demonstrate their declarative knowledge in ways that reveal both the extent and organization of their knowledge. For example, students can group cards with concept words in order of close association. Clusters of closely associated concepts can then be arranged in superordinate and subordinate relationships. In this way students may reveal some of the changes in the structure of their knowledge as they become more expert in a particular knowledge domain (see Royer et al., 1993, for a more extensive discussion of the assessment of knowledge representation).

Accuracy

Accuracy refers to the level of error or distortion in the information generated in assessment practice. Accuracy concerns include the extent to which the intended data collection procedures were actually implemented. For example, say students were being assessed in terms of critical thinking by being asked to develop solutions for ambiguously described problems. The idea is to have students demonstrate cognitive flexibility in defining the problem and developing a number of plausible solutions (assume a representative task for critical thinking for this example). Yet, if in administering the assessment task either in the wording of the assignment or in teacher comments, students are

led to believe they are to derive single, best solutions for a common problem, then accuracy problems arise. The information generated by this assessment practice does not relate to cognitive flexibility per se but rather to identifying a standard problem and deriving a single best solution, and so is inaccurate to a degree. A more accurate assessment practice would result in students developing problem definitions and a range of solutions.

Another accuracy concern is the degree to which assessment practices in the classroom sample the characteristic performance of a given student. As an example of a situation with accuracy concerns, consider the classic test anxiety case. If a student is being assessed in the area of, say, critical thinking and perceives the situation to be one of high anxiety, then not only is thinking critically entering into the assessment information, but also the student's ability to cope with the anxiety. The information then is distorted; it is not accurately reflective of critical thinking.

Accuracy, then, refers to the veracity of the information collected and precision of any subsequent manipulations. It is important to note that accuracy is subsumed under representativeness; that is, each piece (or set) of collected information should be an errorless estimate of the larger assessment target (e.g., can the student add any combination of single-digit numbers?). There will be variations in the information collected to ensure it is representative; thus any variation in student performance or product should be directly related to underlying achievement for the two senses of learning what and for learning how. Student responses, activities, or products should be genuine expressions of achievement of learning goals of interest. For example, in responding to a multiple-choice item on a test of problem solving skills, a student should solve a problem in order to respond rather than simply recall a correct answer. In other words, the student is doing the activities considered to be essential indicators of the achievement of interest. In observing student communication, for example, an accuracy consideration would be seeing two students in oral communication as cooperative dialogue if that is what is going on, or as a threat to obtain free tickets to the new Sylvester Stallone movie if that is what is going on. Thus accuracy includes a determination that there is a match between estimate and target. In addition, accuracy is important in a second sense. Whenever information is collected, it should be recorded exactly, precisely, and unambiguously.

Accuracy considerations, then, follow from establishing a representative mapping of achievement related to tasks and activities. Each component of classroom assessment should be based on a solid information base that is as free from errors as possible in information collection, recording, interpretation, and reporting. For example, if the assessment question is centered on *What is typical student performance?* inaccurate information can result when only best products are selected, as they provide an incomplete and therefore inaccurate representation of the range of work typically completed. Products that might be more typical or that suggest the need for improvement will be missed if only best effort is used when collecting information (Bachor, 1993).

Consistency

Anderson and Bachor (1992) proposed that assessment procedures should be used systematically and consistently so that each student is evaluated for a

given learning outcome on the same basis as any other student. In other words, given the same underlying level(s) of goal attainment, the same kinds of evaluation will be made. In working toward this end, there should be consistency in the quality of procedures for information collection, interpretation, and reporting from one educational area to another and from one student to another *when the goal of assessment is to compare individuals or groups of students against some established criteria.*

Further, each area of schooling should be evaluated with similar standards of information collection and interpretation. For example, the quality and consistency of information collected to evaluate reading attainment should be similar to the quality and consistency of the information collected to evaluate attainment in mathematics or the goal area of physical development.

Several teachers (Bachor & Anderson, 1993) have argued against the use of consistent procedures, claiming that using the same type of procedures across students can result in unfair practices. The rationale for this assertion is that not all students can display what they are learning in the same manner. For example, some students may communicate effectively in writing, whereas others may be more effective visually or orally. As another example, setting a short-answer and essay unit test is unfair if the classroom contains a child who is physically handicapped to the extent that he or she is unable to compose written answers or unable to write at the pace required when taking such a test. Thus any consideration of consistency should be tempered to ensure that any adopted information collection procedure is feasible for all students. However, variations in data collection format can significantly affect the nature of the information collected and the meaningfulness of the conclusions reached.

Generally, however, adopting consistent procedures is suitable, and maintaining consistency increases the probability that representative and accurate estimates are obtained. That is, the goal is to apply the same standards of accuracy and representativeness for each student. An example of a consistency consideration would be that similar observations are being made in each of the observation sessions for that individual student, so that the obtained information can be used to judge if students have met widely held expectations for learning outcomes. Variation in the observed performance is due to variation in the levels of student achievement, not to variation in the methods of data collection.

In order to consider the utility of the criteria of representativeness, accuracy, and consistency in the evaluation of classroom assessment practice, two currently popular forms of assessment practice, observation and portfolios, are described. The use of these practices is discussed in relation to the three criteria discussed above: representativeness, accuracy, and consistency.

Information Collection: Observation and Portfolios

Observation and portfolios are two means of collecting information about student achievement that, as noted in the introduction, are considered authentic forms of classroom assessment practice in that the information collected and the way it is collected are considered to be fully compatible with the kinds of activities students do in classroom learning. In this sense observation and

portfolios could be viewed as formats well adapted to enhance the representativeness of assessment practice.

Observation

Observation typically is used to collect information about the behavior of students (e.g., the extent to which they cooperate or keep the classroom rules) and to note their activities as they learn (e.g., the degree to which they exchange ideas or present a viewpoint). It appears that the use of observation as an assessment practice is increasing in the elementary classroom (Bachor & Anderson, 1990; Board of Education of the City of Etobicoke Writing Committee, 1987; British Columbia Ministry of Education, 1989, 1990a, 1990b; Saskatchewan Education, 1991) and its use is relatively unstructured (Nicholson & Anderson, 1993). For example, the BC Ministry of Education (1989) introduced the use of observation as an assessment procedure in its draft version of the *Primary Program*:

> Observation is an important method of evaluating a variety of learning that cannot be assessed in other ways. For example, interpersonal behaviour, feelings, performance, thinking processes while solving problems, curiosity, and creativity are best noted through observation. Observation can help a teacher determine a pupil's knowledge of content, degree of skill development, and ability to apply skills. By observing pupils systematically in the natural setting of the classroom, the teacher can identify each child's unique interests, personality, learning style, strengths, differences, and difficulties. The teacher can use this information to plan programs that best meet the needs of individual pupils in the class. (p. 11.1)

We argue that using observation in the manner recommended above should be accompanied by close attention to the three criteria of representativeness, accuracy, and consistency.

The use of observation, as with every classroom assessment practice, should require that the teacher determine and articulate the goals of learning and the kinds of student performance or behavior associated with the achievement of these goals. In addition, observation requires that the teacher determine the situation(s) in which this performance is likely to occur.

As noted, a main rationale for the use of observation is that it is an authentic form of assessment because the kinds of things students are doing are directly related to learning activities, and hence the observational techniques are not decontextualized. The practice can be viewed as being representative in that the students' performances are closely related to learning outcomes.

Another representativeness consideration is the sampling of student activities. Sufficient numbers of observations each of adequate duration should be made in order to ensure that the activities observed were a representative sampling of the student's performance. What is sufficient would depend on the nature of the activities under consideration in that simple and relatively stable activities such as any discrete achievements in building vocabulary or increasing computational skills would require fewer observations than more complex activities such as critical thinking or more latent characteristics such as self-esteem or creativity. In terms of duration, an observation period would have to be long enough to allow for the targeted performance to occur, and for

the performance to be completed or at least enough seen that the gist of the performance is captured in the observation.

Accuracy in observation-based assessment concerns the veracity of the observations themselves and the procedures used to record them. Veracity of observations would be difficult to check in a classroom situation. In a more controlled situation multiple observers and perhaps videorecording events could be used to investigate this aspect of accuracy. Nevertheless, in classroom situations the veracity of observations can be enhanced by describing the foci of observations in low inference terms. This minimizes the extent to which observations are evaluated as they are being made. When possible and feasible, it is also helpful to develop and use a structured observation guide.

The consistency of observation procedures is also important. Observational procedures should be applied unvaryingly across students and for the same student across time. The first aspect of observational consistency is directed at ensuring students are observed in a similar manner and that approximately the same number of behavior samples are gathered for each student. In the second, consistency refers to collecting a similar number of samples across activities around the same period of time so that, as much as possible, each activity is portrayed in a parallel manner.

Portfolios

Portfolios are collections of samples of student work that are intended to represent the achievement of particular learning goals. Portfolios are described by the BC Ministry of Education (1992) as being collections

> of student work, gathered by the teacher or student over time, provide a rich array of information and evidence of learning ... When students and teachers develop portfolios that consist of judicious sampling of work done, however, collections become an integral part of learning. It is the substance of the portfolio that is important, not the form ... A portfolio must contain evidence of reflection and self-evaluation as this represents the student's efforts to learn about learning. Reflection and self-evaluation offer a concrete way for students to learn to value their own work, and by extension, to value themselves as learners. (pp. 106-107)

The portfolio generally includes work completed over a period of time and may include repeated drafts of the same piece of work in order to demonstrate the development (improvement) the student has achieved over time. The selection of materials to include in the portfolio generally is done by the student, either independently or with the teacher's assistance in judging whether the selection should be included in the portfolio. As a result of the student's active involvement in building a portfolio, and in contrast to observation and many other forms of assessment practice, in order to make meaningful selection and later evaluative-review decisions, the student should be aware of the learning goals germane to the relevant programs of study. Further, the student must comprehend the purpose of the portfolio. In other words, the teacher is not alone in the control of the assessment practice; the student is not simply responding to discrete items or test questions that have been developed or acquired by the teacher; rather, the student is responsible for assembling a collection of work completed during instruction and learning

that he or she considers to be representative of achievement of the goals of instruction.

The student is also responsible for the development or structuring of the assessment. This development would also involve collaboration between teacher and students on the goals of learning, the criteria for selection of work samples, presentation criteria and formats, and methods and criteria for evaluating the material. Each of the steps of collaboration must be considered in terms of representativeness, accuracy, and consistency.

Representative considerations for portfolios are similar to those for observation: work samples collected should be representative of the kind of performance associated with the achievement of learning goals; and there should be a sufficient number of these work samples taken over the time period under consideration. This requirement should be met by both the student and the teacher. This may require significant collaboration of the teacher and the student to establish goals and relevant (to the individual student and to the class in general) indicators of achievement. The purpose of compiling any portfolio also must be clearly understood by students and the teacher because this purpose will affect the number of samples included and the time period sampled. In addition, the range of material included will need to be considered. If the purpose is to present the full array of activities the student engages in, then work in progress, typical efforts, and best work should be sought. Conversely, if the intent is to illustrate the students' exemplary efforts to show to parents, then only the best material should be sampled. Further, the range of material needs to be considered from the perspective of the students' current capabilities so that written products, audiotapes, and videotapes of activities could be included to ensure that the portfolio includes options to allow the illustration of kind of performance matched with the achievement of learning goals.

Accuracy would focus on whether the targeted data (work samples) are in fact collected. The work samples would have to be the genuine article in the sense the student did indeed produce the work in an instructionally relevant context. For example, issues such as plagiarism and external assistance may come into play, but this would probably not be a major source of problems with the portfolio in the context of classroom assessment. Such issues could be more important where computers are being used, especially when databases are available where material can be readily exchanged. In this case, establishing who wrote what will become difficult. The net result will be that to determine accuracy, that is, a genuine article, could be extremely difficult if not impossible.

Perhaps a greater threat to the accuracy of work samples is found when classroom tasks are completed in groups. Several researchers (Corno & Mandinach, 1983; Marx & Walsh, 1988; Rohrkemper & Corno, 1988) have found that group work provides opportunities for some students to disengage from learning. For example, students may adopt purely management roles in group work. These students, called resource managers by Corno and her colleagues, are adept at gathering and presenting information but leave much of the cognitively engaging work of integrating information to others. In a similar vein, other students may rely almost exclusively on the efforts of others. These

recipient learners (or perhaps nonlearners) acquire and transform information only to glean the gist of what was accomplished in the group. In still other cases, disengagement from participating in the group activity can mask lack of confidence or competence that may be missed by the teacher when group work is prevalent.

Whether students adopt a task management, recipient learner, or other approach to group work is perhaps of lesser importance than is the way that the accuracy of work samples may suffer as a result of different styles of task engagement. Students can produce group work samples that may give the appearance of accurately reflecting learning, but in fact reflect little genuine cognitive effort. Without taking more direct data on the thinking of students as they complete group tasks, the accuracy of the meaning of the work samples may well be reduced.

Several techniques may be used to assess whether work samples completed by students accurately reflect their thinking. Paralleling closely the stimulated recall procedures used in the psychology laboratory, teachers can present work samples to students and ask them to recall the procedures, strategies, and general thinking that created the product. Consider, for example an assessment interview that might be conducted following instruction on poetic techniques. During the interview, a student might be asked to recite a poem she or he has just written. At various points during the recitation, the teacher might interrupt asking why a particular metaphor was used, why onomatopoeia was selected, or why iambic pentameter verse was chosen. By being questioned while the poetry is recited, students' recollection of their thoughts during the initial writing of the poetry are enhanced (Marx et al., 1985). In this way students' reports of their thinking can be stimulated by presenting their academic work during assessment interviews.

Alternatively, students may be instructed to leave evidence as to the way a task was completed during their solution of the problem. Annotated outlines of essays, rough notes of group presentation, and detailed calculations in algebraic problem solving are a few examples of myriad ways students can leave traces of their thinking (Marx et al., 1985). These intermediate academic products can, of course, also be used as the foci of assessment interviews as discussed above.

Consistency considerations can be problematic in the use of portfolios. First, if students are involved in the selection of work samples representative of given learning goals, there must be a common and consistent understanding of what these goals are in a substantive manner between teacher and student and across all participating students. Second, there has to be a common and consistent understanding of the purpose of the portfolio itself, particularly between teacher and student.

Evaluating Evidence and Reporting

The next logical steps in assessing student achievement are to interpret the evidence collected and to report these interpretations. The teacher must make sense of the observations or the portfolio if such information is to inform instruction and aid in making judgments about the achievement of students. In regard to these interpretations, evaluations, and reports, the evidence interpreted should be representative of that collected, and in turn should be repre-

sentative of student performance closely related to learning outcomes. The evaluative judgements should be accurate and consistent. Reporting the evidence and the evaluations should be representative of the student's achievements and the field of learning outcomes, accurate descriptions of evaluations and supporting evidence, and consistent across students in terms of coverage and quality.

Traditionally, teachers have tended to collect information most intensively around reporting periods rather than systematically across the course of the school year. Their typical practice has been to compile any current information and/or to invoke their memories about a student's performance. Following this preparation period, report cards were written and sent home. Parent-teacher interviews followed if they were deemed necessary.

As teachers change evaluative practices, at least two components of the assessment process require clarification. First, the common interpretive practice appears to be to select and report rather than to evaluate the gathered evidence; that is, judgments made by teachers are often quite subjective. Thus we argue that some guidelines are necessary to ensure that the evidence presented is representative (meaningful), accurate, and consistent (fair). Similarly, the rationale and formats of reporting need to be examined to ensure that reporting practices meet the same criteria.

Interpretation

Interpretation of the data collected by assessment practices, such as through observation or portfolio use, should involve some checks of data quality and will general involve substantial reduction of collected information into more manageable aggregates of evidence. To check the quality of the collected data the teacher should evaluate the representativeness, accuracy, and consistency of the gathered information before making any judgments about student achievement. We argue that most of these checks should be done before and during collection of the information. Before examining the resultant data, however, these considerations should be revisited. Further, steps should be taken to ensure that the overall purpose of conducting assessment is maintained.

The reduction of assessment data should be conducted in a way that maintains close linkage to the purpose of the assessment and to the learning outcomes under consideration. To do this will require planning on the part of the teacher. Bachor (1993) recommended that the teacher should begin to interpret information by asking about the extent of evidence required to make judgments about student achievement of a particular learning outcome. This would be directly related to the guidelines we recommend for information collection discussed above. This is recommended in the context of interpreting information as a cautionary step prior to making any substantive decisions about students.

The kind of interpretation teachers make will be one of location: given this information about student performance or product in relation to learning goals, where is the student located in terms of achievement? (Bachor & Anderson, 1993). The issue here is to establish where the students are in relation to established learning goals. To do this student performance is compared with expectations. The expectations that are the basis of this comparison are the

performances expected of achieving and nonachieving students. To enhance the consistency of interpretation, these expectations should be explicitly considered and described by the teacher. As an example, if the learning goal for a class is to *effectively analyze news reports in terms of fact or opinion*, the kinds and qualities of products (in this example written reports) that would indicate achievement or nonachievement should be fully described by the teacher and known by the students for pedagogical reasons. This would facilitate consistent teacher decisions about the achievement status of the students generating the assessment date. It should be noted that the development of a full range of expectation descriptions would be ideal in the maintenance of consistency.

To aid in locating students' accomplishments in an achievement space or on a continuum, numerous scales and frameworks might be used. The use of a well-developed scale or framework could also assist in maintaining consistent interpretation. If teachers have a consistent structure on which to place or locate various levels of students' products or performances, teachers' judgments are likely to be more consistent than if the scale or framework is constantly being created anew for each student with changing meaning associated with similar achievement labels. The most common scale is the letter grade scale. Another, recommended by Bachor (1993), is a framework based on four clusters of performance, where any information relevant to the purpose of classroom assessment could be described under the following labels: capable (can do above level of expectation, competent), characteristic (typical of what the student can do), acquiring (what the student is learning), and unknown (at this time the student is unable to do X).

Regardless of the scale used in reporting, the three recommended criteria must be addressed to determine if students have been located accurately and consistently across the established goals. Possible interpretive errors in making this decision are addressed below.

Possible interpretive errors. Nuttall and Goldstein (1984) list four errors in judgment a teacher can make when reading assessment information: (a) halo effects; (b) leniency or severity effects; (c) central tendency errors, which are essentially errors associated with our criterion of accuracy; and (d) sampling information from a restricted range, which is an error associated with representativeness.

Halo effects occur when teachers' judgments are based on belief rather than performance; that is, teachers judge what they expect overall performance to be rather than examining the activities actually completed. In making errors of leniency or severity, the conclusions reached are not warranted by the collected information. The judgment made is that the performance is poorer or better than is actually the case. Leniency effects can occur when a teacher wants a particular student to do well, accepting an assignment or other work as being "good enough" rather than meeting the desired standard. Severity effects are found when too harsh a judgment is made. The third class of error suggested by Nuttall and Goldstein (1984) is the central tendency error. In this case only typical information is collected while extreme performance is ignored. As a consequence, the full range of potential performance is unknown to the teacher, which would be an error if the purpose of assessment were to evaluate the full range of student performance. The last type of error is in

sampling information from a restricted range of possible performances. The restriction is not in terms of the range of performance of a given task by students as in central tendency error, but rather in terms of the kinds of performances considered. The performances observed would reflect a restricted range of learning and teaching conditions. For example, a student who is being judged on mathematics achievement may be classified incorrectly if the work being considered represents performance relying primarily on the student's reading ability (Bachor, 1993).

Reporting

One purpose of classroom assessment (and a legal requirement) is to report to parents. To some degree teachers have little choice as to the number, format, and timing of the issued report, as such matters typically are prescribed. However, teachers still have to address a number of practical questions when they report: How complete is the presented report to be (the extent of coverage of each goal for any student and to a lesser degree the number of goals to be addressed)? Should the report be supplemented by a conference (assuming one is not required), or if only an oral report is needed should it be written instead? If a conference is to be held with parents, who should participate? To some degree teachers in any school also are varying timing of reporting delivery (Bachor & Anderson, 1993). In these cases teachers are asking about the timing of the delivery of any report; should it be before, after, or during a conference with parents?

Reports on student achievement should be representative in terms of representing the range of goals reported and the range of student achievements in the goal areas. In terms of goal coverage, the report should not be confined to a restricted range of goal areas: all relevant and important goal areas should be targeted. In terms of achievement coverage, the full spectrum of student achievements should be reported. As an example of a restricted range of performance being reported, in the early grades in British Columbia (up to about grade 4) there was an emphasis on reporting only those student achievements that could be perceived as a positive reflection of what the student had accomplished. These were the so-called can-do statements of student achievement. In effect, a restricted range of goal areas and student achievements was reported for many students (Bachor & Anderson, 1993). In some cases, can-do reports were summaries of the student's learning history, rather than a reflection of what was being learned in the current year. In neither case would the resultant report be representative of the established classroom goals.

Accuracy would essentially focus on the truthfulness of the reports. The accuracy would be determined by the preceding steps of information collection, interpretation, and evaluation. Given that these preceding steps were representative, accurate, and consistent, the chance of the report distorting the student's current achievement should be minimal.

Written reports vary from comments alone (termed an anecdotal report) to combinations of commentary and letter grades. An anecdotal report is a narrative on student achievement but may also include summary descriptions and evaluations of student abilities and characteristics. The assignment of evaluations and comments to students is based mainly on teachers' tacit understanding of students (Bachor & Anderson, 1993). Similarly, the basis of the

assignment and the form of letter grades varies greatly, making interpretation of this kind of report more difficult for parents and students. So, certainly, the meaning of symbols, phrases, and other characterizations of student achievement has to be fully explained to parents and used consistently across students and across reporting periods to enhance the utility and meaningfulness of classroom assessment practice. An exception will occur when students with special needs and individualized programs are added to any class. In such cases, there might be some degree of inconsistency in the reporting format and content.

Several new forms of reporting are emerging in popular practice in British Columbia schools. Among these, student self-evaluation and various forms of conferencing are prominent. These two reporting procedures may take a variety of forms; nevertheless, the criteria of representativeness, accuracy, and consistency are unvaryingly important across these forms. To discuss these criteria in relation to reporting is a topic for future work.

Summary

We argue that practitioners need to evaluate carefully classroom assessment and reporting practices. The quality of these practices is not well known, although the importance and presence of classroom assessment practice is considerable. In order to assist practitioners in their evaluation of their classroom assessment practice, we suggest the use of three criteria: representativeness, accuracy, and consistency.

References

Anderson, J., & Bachor, D. (1992). What should a classroom testing program look like? The functional factors of an assessment program in primary classrooms. In D.J. Bateson (Ed.), *Classroom testing in Canada* (pp. 59-68). Vancouver, BC: Center for Applied Studies in Evaluation, University of British Columbia.

Bachor, D. (1993). The changing face of assessment. In J. Andrews & J. Lupart (Eds.), *The inclusive classroom: Educating exceptional children* (pp. 293-329). Toronto, ON: Nelson Canada.

Bachor, D., & Anderson, J.O. (1990). *Assessment practices in the primary program: Description of observed practices, functional factors, and recommendations as to some general principles: Final report*. Victoria, BC: Ministry of Education.

Bachor, D.G., & Anderson, J.O. (1993). *Assessment practices in the elementary classroom: Perspectives of stakeholders*. Victoria, BC: Ministry of Education.

Board of Education of the City of Etobicoke Writing Committee. (1987). *Making the grade: Evaluating student progress*. Scarborough, ON: Prentice-Hall.

British Columbia Ministry of Education. (1989). *The primary program: Response draft*. Victoria, BC: Author.

British Columbia Ministry of Education. (1990a). *The primary program: Foundation document*. Victoria, BC: Author.

British Columbia Ministry of Education. (1990b). *The primary program: Resource document*. Victoria, BC: Author.

British Columbia Ministry of Education. (1992). *Intermediate program: Draft, March 1992*. Victoria, BC: Author.

Corno, L., & Mandinach, E.B. (1983). The role of cognitive engagement in classroom learning and motivation. *Educational Psychologist, 18*, 88-108.

Marx, R.W., & Walsh, J. (1988). Learning from academic tasks. *Elementary School Journal, 88*, 207-219.

Marx, R.W., Winne, P.H., & Walsh, J. (1985). Studying student cognition during classroom learning. In M. Pressley & C.J. Brainerd (Eds.), *Cognitive learning and memory in children: Progress in cognitive development research* (pp. 181-203). New York: Springer Verlag.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229-258.

Nicholson, D.J., & Anderson, J.O. (1993). A time and place for observations: Talking with primary teachers about classroom assessment. *Alberta Journal of Educational Research, 39*, 363-374.

Nuttall, D., & Goldstein, H. (1984). Profiles and graded tests: The technical issues. In J. Mortinare (Ed.), *Profiles in action* (pp. 5-150). London: Further Education Unit.

Rohrkemper, M., & Corno, L. (1988). Success and failure on classroom tasks: Adaptive learning and classroom teaching. *Elementary School Journal, 88*, 297-312.

Royer, J.M., Cisero, C.A., & Carlo, M.S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63*, 201-243.

Saskatchewan Education. (1991). *Student evaluation: A teacher handbook*. Regina, SK: Author.

Shepard, L.A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Research in education: Vol. 19* (pp. 405-450). Washington, DC: American Educational Research Association.

Swing, S.R., & Peterson, P. (1988). Elaboration and integrative thought processes in mathematics learning. *Journal of Educational Psychology, 80*, 54-66.

Walsh, J. (1992). Promoting self-regulation in the learning disabled. *Exceptionality Education in Canada, 2*, 1-14.

Walsh, J. (1993). The promise and pitfalls of integrated strategy instruction. *Journal of Learning Disabilities, 26*, 438-442.

Weinstein, C.E., & Meyer, D.K. (1991). Implications of cognitive psychology for testing: Contributions from working in learning strategies. In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition* (pp. 40-61). Englewood Cliffs, NJ: Prentice-Hall.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

Wilson, R.J. (1990). Classroom processes in evaluating student achievement. *Alberta Journal of Educational Research, 36*, 4-17.

For Contributors

Please prepare manuscripts in accordance with these guidelines.

Format

Manuscript in triplicate.
Typewritten or printed, double-spaced throughout.
Abstract of approximately 100 words on a separate sheet.
Manuscript not to exceed 6,500 words excluding graphics.
Author's name and affiliation on a separate title page; title only on first page of manuscript to ensure anonymity in the review process.

Referencing

Sources in parentheses after each reference (direct or otherwise), giving author's name (unless mentioned in text), year of publication, and page number(s) for direct quotes.
Direct quotes of 40 words or less in double quotation marks incorporated into text; those longer than 40 words in indented block format. Page numbers must be given.
All sources listed alphabetically at the end of the manuscript under the heading *References* using the APA style.
Explanatory notes, numbered consecutively and marked in the text with superscript numerals, may appear before the References under the heading *Notes*; any citations in notes follow the same format as other references.

Graphics

Each table and figure on a separate page at the end of the manuscript.
Figures, charts, and diagrams in camera-ready form.
Number tables in Arabic numerals and center over each heading, e.g., *Table 1.*
Number figures in Arabic numerals before the caption at foot of each graph, e.g., *Figure 1. ...*
Indicate placement of figures and tables in text, i.e., *Insert Table 3 here.*

Style

AJER's editorial style conforms closely to the *Publication Manual of the American Psychological Association.*
Editorial changes may be made to manuscripts.
For spelling, consult *Webster's New Collegiate Dictionary;* spelling in quoted material remains as in the original.

Diskettes

If manuscript is accepted for publication please supply a diskette from a Macintosh, IBM, or IBM compatible system. Microsoft Word and WordPerfect are the preferred software.

Submission to *AJER* implies the manuscript is not being considered for publication elsewhere.

AJER is indexed in the *Alberta Education Index, Canadian Education Index, Current Contents/Social and Behavioral Sciences, ERIC/Current Index to Journals in Education,* and *Social Science Citation Index;* appropriate articles are abstracted in *Cognitive Development Abstracts, Educational Administration Abstracts, Educational Technology Abstracts, Family Abstracts, Language Behavior Abstracts, Multicultural Education Abstracts, Psychological Abstracts, Research into Higher Education Abstracts, School Organization and Management Abstracts,* and *Sociology of Education Abstracts,* and reviewed by the *Journal for Research in Mathematics Education.*

If undelivered return to:

The Alberta Journal of Educational Research
4-116 Education North
University of Alberta
Edmonton, AB
T6G 2G5

Publications Mail Registration Number 1436